

THE ROLE OF CONTENT-RICH VISUALS
IN THE L2 ACADEMIC LISTENING
ASSESSMENT CONSTRUCT

By Roman O. Lesnov

A Dissertation

Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy
in Applied Linguistics

Northern Arizona University

May 2018

Approved:

Joan Jamieson, Ph.D., Chair

Soo Jung Youn, Ph.D.

Jesse Egbert, Ph.D.

Ruslan Suvorov, Ph.D.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

ABSTRACT

THE ROLE OF CONTENT-RICH VISUALS IN THE L2 ACADEMIC LISTENING ASSESSMENT CONSTRUCT

ROMAN O. LESNOV

Despite the growing recognition that second language (L2) listening is a skill incorporating the ability to process visual information along with the auditory stimulus, standardized L2 listening assessments have been predominantly operationalizing this language skill as visual-free (Buck, 2001; Kang, Gutierrez Arvizu, Chaipapae, & Lesnov, 2016). This study has attempted to clarify the nature of the L2 academic listening assessment construct regarding the role of visual information.

This goal was achieved by developing an interpretive argument for including video-based visuals in L2 academic listening tests. Particular attention was paid to the role of content-related visuals that provided graphical illustration, description, or explanation of the auditory listening message. Using Kane's validity framework, the explanation inference was of primary concern to this study because it is used to justify the measured construct (Kane, 1992; 2004; 2006; 2013).

The explanation inference was supported by two types of evidence. First, the performances of 143 English as a second language (ESL) and English as a foreign language (EFL) students on an academic English listening comprehension test were quantitatively analyzed for the effect of delivery mode (i.e., audio-only vs video-based) and its relationships with test-takers' listening proficiency (i.e., lower vs higher), item video-dependence (i.e., whether or not an item was cued by video), item type (i.e., local vs global), and viewing behavior (self-reported on a scale from 1-did not watch the video

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

to 5-watched all of the video). Analyses were based on both classical test theory (i.e., ANOVA and correlations) and item response theory (i.e., Rasch analysis). In the video-based version of the test, content-rich videos were used, defined as videos containing relevant graphical content-related visual cues for 60% of the video length.

The findings showed that video-dependent items were easier with videos than without for both lower-level and higher-level test-takers, regardless of item type. Video-independent items were unexpectedly harder with videos in general. In particular, video-independent global items were harder in the video-based mode than in the audio-only mode for lower-level test-takers. Viewing behavior had a weak positive relationship with listening comprehension, regardless of proficiency.

Second, stakeholders' perceptions about using content-rich videos were investigated. Using a questionnaire, the same 143 test-takers provided their perceptions of test difficulty, motivation towards listening, listening authenticity, and whether content-rich videos should be used in high-stakes academic listening tests. The effects of mode and proficiency on these perceptions were examined. Similarly, 310 ESL and EFL teachers provided their opinions about the effects of content-rich videos on listening difficulty, motivation, authenticity, and using content-rich videos in L2 listening tests. The effects of teachers' background (i.e., professional location, education level, and teaching-related experience) on their perceptions were examined.

Test-takers found the video-based mode easier than the audio-only mode; however, their perceptions of motivation, authenticity, and using videos in tests were not affected by mode. Regarding video use perceptions, test-takers were in favor of including content-rich videos in L2 academic listening tests. Teachers were more favorable towards

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

the video-based mode than the audio-only mode in terms of listening difficulty, motivation, authenticity, and using videos in L2 academic listening tests.

The study has discussed how these findings supported the interpretive argument for including content-rich video-based visual information into the assessment construct of L2 academic listening comprehension. Challenges revealed by the findings were also addressed, with limitations acknowledged. The study also offered theoretical and practical implications for the field of L2 assessment. As its primary implication, the study recommends that test developers start using content-rich visual information in L2 academic listening tests.

Keywords: academic, assessment, difficulty, listening, perceptions, stakeholders, test, validity, video, visual aids

Acknowledgements

I dedicate this dissertation project to my wife, Nina Liesnova, who has supported me immensely throughout my Ph.D. journey. I thank you for your unconditional love, your patience, and your unwavering faith in me. I love my project and my career, but it is nothing compared to how much I love you.

I would also like to acknowledge my advising professor, Dr. Joan Jamieson, for guiding and supporting me through this dissertation project. You have helped me to develop as a researcher, all the while being an endless source of encouragement, support, and inspiration. Thank you for many long hours you spent reading my dissertation and writing your invaluable notes all over the paper. It would be hard to become as great of a scholar and teacher as you are, but, looking at you, I can't help but follow your steps.

I am grateful to the members of my dissertation committee, Dr. Soo Jung Youn, Dr. Ruslan Suvorov, and Dr. Jesse Egbert. Without your feedback, it would have been much harder, if not impossible, to complete this dissertation. I hope this project has been beneficial and interesting for you.

I am also thankful to everybody who assisted me in completing this dissertation, including my former and current students, colleagues, friends, and relatives in the US, Russia, Ukraine, and the rest of the world. Thank you for making this project happen! Special thanks go to Dr. Maria Nelly Gutierrez Arvizu, Ms. Julia Shut, Dr. Stanley Van Horn, and the leadership of online English language schools White Rabbit and EnglishDom for their help with data collection.

Finally, I am happy to acknowledge the following organizations for supporting this project financially: Paragon Testing Enterprises Inc., the British Council, the

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

National Federation of Modern Language Teachers' Associations, the Educational Testing Service, and the Northern Arizona University.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Table of Contents

ABSTRACT	ii
Acknowledgements	v
Chapter 1. Introduction	1
Problem Statement	3
Purpose of the Study	4
Research Questions	4
Overview of Method	5
Participants	5
Instruments and procedures.	5
Significance of the Study	6
Role of the Researcher	7
Assumptions	8
Delimitations	8
Definition of Key Terms	9
Organization of the Dissertation	10
Chapter 2. Literature Review	12
Defining and Justifying an Assessment Construct.....	13
Construct definition	13
Construct justification	14
Summary	22
Organization of Review	23
Theoretical Understanding of L2 Listening Skill	24

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Theoretical model of L2 listening.....	24
Input decoding.....	26
Lexical search.....	26
Syntactic parsing.....	27
Meaning construction.....	27
Discourse construction.....	28
Summary of the model.....	28
Visual information in L2 listening.....	29
Types of visual information.....	29
Theorized benefits of visual information.....	32
Factors affecting the role of visual information.....	36
Proficiency.....	36
Visual literacy.....	36
Viewing behavior.....	37
Evolution of L2 listening definitions.....	38
Summary.....	41
TLU Domain: L2 Academic Listening.....	41
Contextualized approach to academic listening.....	41
Features of academic discourse.....	43
Structural features.....	44
Stylistic features.....	45
Linguistic features.....	45
Visual features.....	46

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Academic listening comprehension	47
Taxonomies of academic listening sub-skills	47
Evaluation of lecture comprehension.....	48
Summary.....	50
Comparative Empirical Studies	50
Video effect.....	51
Positive video effect.....	51
Negative video effect.....	56
Neutral video effect.....	58
Reasons for mixed findings.....	61
Video type.....	62
Viewing behavior.....	64
Item video-dependence.....	65
Video effect at the item level.....	66
Summary.....	68
Stakeholders' Perceptions.....	68
Authenticity.....	69
Difficulty.....	70
Motivation.....	71
Summary.....	72
Research Gaps.....	72
Research Questions and Hypotheses	74
Chapter 3. Method	77

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Purpose of the Study	77
Participants.....	78
ESL/EFL learners.....	78
ESL/EFL teachers.....	80
Measures	83
Academic listening comprehension test.....	83
Search for authentic video lectures.....	84
Video recording of lectures.....	85
Item development.....	92
Test prototyping.....	94
Test piloting	95
Empirical confirmation of items' visual-related designs.....	97
Anchor test.....	99
Test-takers' questionnaire.....	101
Teachers' questionnaire	103
Procedures.....	106
ESL/EFL learners.....	106
ESL/EFL teachers.....	109
Research Design.....	110
Variables in the Study.....	111
Data Analysis	113
Preliminary analyses.....	113
Descriptive statistics.....	114

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Psychometric properties.....	114
Determination of group equivalence.....	115
Operationalization of test-takers' proficiency.	116
Research question 1.	116
Research question 1.1.	116
Classical analyses.....	117
Rasch analyses.	119
Research question 1.2.	121
Research question 1.3.	123
Summary.....	124
Research question 2.	124
Research question 2.1.	124
Research question 2.2.	126
Summary.....	127
Alpha level and effect sizes.	127
Chapter 4. Results.....	129
Data Screening.....	129
Data quality.....	130
Data accuracy.....	131
Missing data.....	132
Outliers.....	132
Preliminary Analysis for the Anchor Test.....	132
Psychometric properties.....	133

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Descriptive statistics.....	133
Item performance statistics.....	134
Reliability analyses.....	137
Summary.....	137
Determination of group equivalence.....	138
Operationalization of test-takers' proficiency.....	138
Preliminary Analyses for the ALC Test.....	139
Psychometric properties.....	139
Descriptive statistics.....	139
Item performance statistics.....	140
Reliability analyses.....	144
Summary.....	146
Data subset connection for Rasch analysis.....	146
Results for Research Question 1.....	148
Research question 1.1: Classical Analysis.....	149
Assumption check.....	149
ANOVA #1 on all items.....	150
ANOVA #2 on video-dependent items.....	151
ANOVA #3 on video-dependent local items.....	153
ANOVA #4 on video-dependent global items.....	154
ANOVA #5 on video-independent items.....	155
ANOVA #6 on video-independent local items.....	157
ANOVA #7 on video-independent global items.....	158

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Summary	159
Research question 1.1: Rasch Analysis.	159
Delivery mode.....	160
Proficiency and mode.	161
Video-dependence and mode.....	161
Video-dependence, proficiency, and mode.....	162
Video-dependence, item type, and mode.....	163
Video-dependence, proficiency, item type, and mode.....	163
Summary.....	164
Research question 1.2.	165
Delivery mode.....	166
Mode and proficiency.	167
Lower proficiency.	167
Higher proficiency	169
Comparison across proficiency levels	170
Summary.....	171
Research question 1.3.	172
Research Question 2	175
Research question 2.1.	175
Operationalization of variables.....	175
Assumption check.....	178
Difficulty perceptions.	178
Motivation perceptions.	180

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Authenticity perceptions.....	181
Video use perceptions.....	182
Summary.....	184
Research question 2.2.....	184
Operationalization of the variables.....	185
Assumption check.....	187
Difficulty perceptions.....	187
Motivation perceptions.....	189
Authenticity perceptions.....	191
Video use perceptions.....	192
Summary.....	193
Chapter 5. Discussion.....	194
Summary of Findings.....	195
Test difficulty.....	195
Stakeholders' perceptions.....	196
Listening Comprehension Difficulty.....	197
Items with video-dependent design.....	197
Items with video-independent design.....	201
Overall test difficulty.....	202
Viewing behavior.....	203
Stakeholders' Perceptions.....	204
Test-takers' perceptions.....	204
Teachers' perceptions.....	206

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

The Interpretive Argument	208
Test domain.....	208
Evaluation inference.	209
Generalization inference.	210
Explanation inference.	211
Summary.....	212
Implications.....	215
Theoretical implications.....	215
Methodological implications.	216
Assessment implications.....	221
Pedagogical implications.	225
Limitations of the Study.....	227
Directions for Future Research	228
Conclusion	231
References.....	232
Appendix A1. Sampling Frame for TESOL-Affiliated Organizations.....	254
Appendix A2. TESOL Affiliates Selected for the Study.....	257
Appendix B. Video Listening Passages Found on the Internet	260
Appendix C1. Recording Instructions.....	261
Appendix C2. Visual Configurations of the Four ALC Test Videos.....	261
Appendix D. ALC Test: Consent, scripts, items, specifications.....	271
Appendix E. Item Video-Dependence Survey.....	286
Appendix F. Classifying ALC Items as Video-Dependent vs. Video-Independent	317

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Appendix G. YouTube Video Listening Passages for the Anchor Test	318
Appendix H. Anchor Listening Test (scripts, items, table of specifications).....	319
Appendix I. Test-takers' Questionnaire	326
Appendix J. Teachers' Questionnaire	331
Appendix K. Rasch Analysis Specifications Template	338
Appendix L. <i>Post hoc</i> Comparisons for Education-Experience Interaction (RQ 2.2)....	339

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

List of Tables

Table 2.1. Inferences in Kane’s Interpretive and Validity Arguments for an L2 Test	17
Table 2.2. Historical Overview of L2 Listening Definitions	39
Table 2.3. Positive Video Effect on Comprehension: Comparative Studies	53
Table 2.4. Negative Video Effect on Comprehension: Comparative Studies.....	57
Table 2.5. Neutral Video Effect on Comprehension: Comparative Studies	59
Table 3.1. Participants’ Affiliations with Language Schools or Online Platforms	79
Table 3.2. Learners’ Location, Age, and Gender.....	80
Table 3.3. Learners’ Native Languages by Location.....	80
Table 3.4. Teachers’ Demographics	82
Table 3.5. Teachers’ Native Languages by Region	82
Table 3.6. Listening Content Criteria for Selecting YouTube Video Passages.....	84
Table 3.7. Characteristics of the Listening Passages	88
Table 3.8. Word Counts in Each Video’s Pictures and Graphs	92
Table 3.9. Final Classifications of the ALC Test Items by Video-Dependence.....	99
Table 3.10. Features of the Anchor Listening Passages	100
Table 3.11. Test-takers’ Questionnaire Design.....	102
Table 3.12. Initial Teachers’ Questionnaire Design and Reliability.....	104
Table 3.13. Revised Teachers’ Questionnaire Design and Reliability	105
Table 3.14. Recruitment of Learners	107
Table 3.15. Dependent Variables in the Study	112
Table 3.16. Independent Variables in the Study.....	113
Table 3.17. Expected Psychometric Properties for the Anchor and ALC Tests.....	115

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Table 3.18. Collection of ANOVAs for Research Question 1.1.....	118
Table 4.1. Descriptive Statistics for Anchor Test Scores	134
Table 4.2. Anchor Test Items' Properties	135
Table 4.3. Anchor Test: Summary Statistics for Rasch Test-Takers and Items Facets ..	137
Table 4.4. Descriptive Statistics for the ALC Test Score	140
Table 4.5. Item-Level Psychometric Properties of the ALC Test Items.....	141
Table 4.6. ALC Item Statistics by Testlet, Video-Dependence, and Item Type.....	143
Table 4.7. ALC Test: Summary Statistics for Rasch Test-Takers and Items Facets	145
Table 4.8. Comparison of Rasch Properties for ALC Items between Two Samples.....	148
Table 4.9. Collection of ANOVAs for Research Question 1.1.....	149
Table 4.10. Descriptive Statistics for 24 ALC Test Items by Mode and Proficiency.....	151
Table 4.11. Results of Two-Way Factorial ANOVA for RQ 1.1: Overall ALC Scores	151
Table 4.12. Descriptive Statistics for 14 Video-Dependent Items.....	152
Table 4.13. Two-Way Factorial ANOVA on 14 Video-Dependent Items	152
Table 4.14. Descriptive Statistics for 7 Video-Dependent Local Items	153
Table 4.15. Two-Way Factorial ANOVA on 7 Video-Dependent Local Items	154
Table 4.16. Descriptive Statistics for 7 Video-Dependent Global Items.....	154
Table 4.17. Two-Way Factorial ANOVA on 7 Video-Dependent Global Items	155
Table 4.18. Descriptive Statistics for 10 Video-Independent Items	156
Table 4.19. Two-Way Factorial ANOVA on 10 Video-Independent Items.....	156
Table 4.20. Descriptive Statistics for 5 Video-Independent Local Items	157
Table 4.21. Two-Way Factorial ANOVA on 5 Video-Independent Local Items.....	157
Table 4.22. Descriptive Statistics for 5 Video-Independent Global Items	158

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Table 4.23. Two-Way Factorial ANOVA on 5 Video-Independent Global Items.....	159
Table 4.24. Rasch Measurement Report for Delivery Mode.....	160
Table 4.25. Rasch Interaction: Proficiency in the Mode Effect.....	161
Table 4.26. Rasch Interaction: Video-Dependence in the Mode Effect.....	162
Table 4.27. Rasch Interaction: Video-Dependence and Proficiency for Mode.....	162
Table 4.28. Rasch Interaction: Video-Dependence and Item Type for Mode.....	163
Table 4.29. Rasch Interaction: Video-Dependence, Proficiency, Item Type for Mode .	164
Table 4.30. Comparison of Classical and Rasch Analyses Results for RQ 1.1.....	165
Table 4.31. ALC Test Items: Rasch Item-Mode Interactions.....	167
Table 4.32. Effects of Mode on Individual Items' Difficulties at Lower Proficiency....	168
Table 4.33. Effects of Mode on Individual Items' Difficulties at Higher Proficiency ...	169
Table 4.34. Comparison of ALC Items' Difficulty in the Video-Based Mode.....	171
Table 4.35. Spearman's Correlations for Viewing Behavior Ratings by Testlet.....	173
Table 4.36. Descriptive Statistics for Viewing Behavior by Proficiency and Testlet	173
Table 4.37. Correlation Analyses for Viewing Behavior and ALC Test Scores.....	174
Table 4.38. Correlations for Difficulty, Motivation, and Authenticity Perceptions.....	176
Table 4.39. Reliabilities for Difficulty, Motivation, and Authenticity Perceptions.....	177
Table 4.40. Descriptive Statistics for Test-Takers' Difficulty Perceptions.....	179
Table 4.41. Two-Way Factorial ANOVA on Test-Takers' Difficulty Perceptions.....	179
Table 4.42. Descriptive Statistics for Test-Takers' Motivation Perceptions.....	180
Table 4.43. Two-Way Factorial ANOVA on Test-Takers' Motivation Perceptions.....	181
Table 4.44. Descriptive Statistics for Test-Takers' Authenticity Perceptions.....	182
Table 4.45. Two-Way Factorial ANOVA on Test-Takers' Authenticity Perceptions ...	182

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Table 4.46. Descriptive Statistics for Test-Takers' Video Use Perceptions.....	183
Table 4.47. Two-Way Factorial ANOVA on Test-Takers' Video Use Perceptions	184
Table 4.48. Internal Consistency Reliabilities for Teachers' Perceptions	186
Table 4.49. Operationalizations of the Independent Variables for RQ 2.2.....	186
Table 4.50. Descriptive Statistics for Teachers' Perceptions on Difficulty.....	188
Table 4.51. Three-Way Factorial ANOVA on Teachers' Perceptions on Difficulty	189
Table 4.52. Descriptive Statistics for Teachers' Perceptions on Motivation.....	190
Table 4.53. Three-Way Factorial ANOVA on Teachers' Perceptions on Motivation ...	190
Table 4.54. Descriptive Statistics for Teachers' Perceptions on Authenticity	191
Table 4.55. Three-Way Factorial ANOVA on Teachers' Perceptions on Authenticity .	192
Table 4.56. Descriptive Statistics for Teachers' Perceptions on Video Use	192
Table 4.57. Three-Way Factorial ANOVA on Teachers' Perceptions on Video Use	193
Table 5.1. Interpretive Argument for the Video-Based ALC Test	214
Table F.1. Collective Data for Items with Originally Video-Dependent Design	317
Table F.2. Collective Data for Items with Originally Video-Independent Design.....	317

List of Figures

Figure 2.1. Organization of the literature review.....	24
Figure 3.1. A screenshot from the content-rich video for the Taxes lecture.	90
Figure 3.2. Illustration for the investigation of test-takers' behavior.	111
Figure 4.1. The Wright map for the anchor test.....	136
Figure 4.2. The Wright map for the ALC test items.....	144

Chapter 1

Introduction

The rapid rise of technology in the 21st century has dramatically altered the way students are taught around the world. The widespread access to and growing ability to use technology-enhanced presentations, audio-video equipment, and interactive distance learning platforms have fostered the use of visual aids in educational contexts. The visual element has established itself as an integral part of lecture-oriented classes in various academic disciplines (Collis & Wende, 2002; Lynch, 2011). In addition to seeing professors' body language, learners nowadays are likely to be exposed to visuals related to the content of a lecture, such as an interactive PowerPoint presentation displaying graphical and textual information (Lynch, 2011). As a result, a successful comprehension of a lecture today may not only depend on the understanding of an auditory stimulus but may also largely rely on the ability to interpret content-related visual aspects of the lecture. This highlights the issue of whether L2 academic listening should be considered a visual-inclusive skill.

The field of second and foreign language (L2) assessment has been actively debating the issue of including visual processing as part of the listening ability. Many studies showed that the inclusion of visual information could make an L2 listening construct more authentic (e.g., Ockey, 2007; Wagner, 2010a; 2013). This view is in line with contemporary conceptualizations of the listening skill, which largely acknowledge the role of visual information for successful listening (e.g., Field, 2008; Flowerdew and Miller, 2005; Richards, 1983). In addition to recognizing the importance of non-verbal cues (e.g., eye expression), scholars nowadays have started gravitating towards

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

considering visuals that are content-related (e.g., visual aids in a lecture) as part of listening (e.g., Rost, 2016). Despite this, L2 listening tests developers continue to favor “pure” definitions of listening, which view visual information as a source of construct-irrelevant variance (e.g., Buck, 2001; Chastain, 1976; Lado, 1964). High-stakes listening assessments today tend to shy away from including visuals of any kind, with a rare exception of using pictures for motivational purposes (Kang, Gutierrez Arvizu, Chaipupae, & Lesnov, 2016). Taking into account the widespread availability of video technology, which is thought to be most capable of reflecting the visual reality of L2 contexts and no longer technologically problematic, the unwillingness to build video-based listening assessments remains an unsolved mystery (Gruba, 2014; Li, 2013).

In light of this, justifying the use of a visually-inclusive L2 listening assessment construct seems of primary importance. For L2 *academic* listening assessments, such a construct would be expected to include content-related visuals along with non-verbal cues and situational visuals. This would properly reflect authentic university contexts in terms of the accessibility and kinds of visual information. The ratification of such a construct requires empirical evidence supporting the inclusion of visual information. This evidence could come from several sources, two of which are test-takers’ performance on an L2 listening test and stakeholders’ perceptions (Bachman and Palmer, 2010; Chapelle, Enright, & Jamieson, 2008; Gruba, 2014). L2 test-takers’ systematic contrastive performance on a video-based listening test versus an audio-only counterpart would indicate a change in a measured construct. L2 stakeholders’ (e.g., test-takers’ and teachers’) perceptions about the role of visual information in the listening skill and

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

listening assessments could either support or oppose the visual-inclusive nature of a listening construct.

Problem Statement

Although the L2 listening skill is generally believed to incorporate the ability to interpret visual cues along with the auditory message, the L2 academic listening assessment construct is still viewed as visual-free by some theorists and operationalized as audio-centric by the majority of high-stakes test developers. In these respects, the assessment construct of academic listening comprehension remains underrepresented in terms of the role of visual information that is most typical of academic contexts. Empirical evidence is needed that would support the inclusion of content-related visuals in the L2 academic listening construct.

Attempts to obtain evidence for a visually-inclusive construct have been either inconclusive or missing. First, studies investigating the effects of videos on L2 listening comprehension have produced inconclusive results. This may be attributed to different video types used in the studies, which could have affected test-takers' viewing behavior and performance. Previous attempts to classify videos into context versus content types have been mostly unsuccessful (Ginther, 2002; Suvorov, 2015a). A more meaningful classification was required that could control for the extent to which videos are rich in content-related visuals and helpful for understanding the listening message. No attempts have been made to investigate the degree to which comprehension items in a listening test could be keyed from the content-related cues (i.e., item video-dependence) and how this degree affects comprehension of items.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Second, research into stakeholders' perceptions about the role of content-rich visual information in academic listening is scarce. Few studies investigated test-takers' perceptions about helpfulness of different video types for listening comprehension. Although research into L2 test-takers' opinions about the effects of video-based visual information is substantial, it has largely failed to control for video type. Moreover, there is a dearth of research into L2 educators' perceptions about the nature of the L2 academic listening construct as well as about whether video-based content-rich visual information should be used in L2 listening comprehension tests.

Purpose of the Study

Using Kane's argument-based validity framework (e.g., Kane, 2006; 2013), this study aimed to develop an interpretive argument for including content-related visual information into the assessment construct of L2 academic listening comprehension. This purpose was primarily achieved by investigating (1) L2 students' performance on an academic listening test and (2) L2 learners' and teachers' perceptions about helpfulness of content-rich videos for comprehension and the use of such videos in listening tests.

Research Questions

The study was guided by two major research questions.

1. Do content-rich videos affect L2 academic listening comprehension difficulty? An affirmative answer for this question was expected. This would signal a difference in the nature of the academic listening construct and support the inclusion of content-related visuals into the construct. This question was investigated with fuller methodological rigor compared to previous studies. Specifically, the study aimed for clearer definitions and closer control for video type and item video-dependence.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

2. Do stakeholders' perceptions lend support for using content-rich videos in the L2 academic listening assessment construct? Higher-level L2 learners and teachers were expected to have positive opinions about using content-rich videos in listening tests while lower-level learners might be unhappy with viewing visuals during listening tests. These effects would further advance the argument for considering the processing of content-rich visuals as part of the L2 academic listening construct.

Overview of Method

Participants. The sample of participants consisted of two main groups – learners and teachers. To sample learners, adult ESL/EFL learners from different USA-based and foreign schools were invited to take an online listening assessment consisting of an academic listening test, a listening proficiency test (henceforth, the anchor test), and a questionnaire. Each learner was randomly assigned to either the audio-only or video-based version of the test and the questionnaire. The number of participants was 143.

To sample teachers, organizations affiliated with the Teaching English as a Second Language (TESOL) International Association were randomly selected from the list of worldwide TESOL affiliates and invited to participate in the study. The number of consented respondents was 310.

Instruments and procedures. To answer the first research question, three instruments were developed, namely the academic listening comprehension test (ALC test; 4 lectures, 24 multiple-choice items), the anchor test (2 lectures, 12 multiple-choice items), and test-takers' questionnaire (10-12 questions depending on the version). The ALC test and test-takers' questionnaire had two versions, audio-only and video-based. The video-based version of the ALC test used content-rich videos, defined as videos

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

displaying content-related visual cues, which are semantically congruent with the auditory stimulus, for about 60% of the video length (a more thorough definition is provided in the Key Terminology section). Test performances of the audio-only group and the video-based group were compared at both the test level and the item level, using the combination of classical test theory analyses, such as ANOVA, and Rasch analysis. The role of test-takers' listening proficiency in the mode effect was also investigated. The anchor test measured a construct of visual-free academic English listening comprehension. It was used solely to estimate test-takers' listening proficiency. Proficiency was operationalized based on test-takers' anchor test score and served as an independent variable. Another independent variable was item video-dependence. It was used to label each individual ALC test item as either video-dependent or video-independent. The presence of video-dependence for each item in the ALC test was determined based on ESL/EFL teachers' and learners' judgements and performances.

To answer the second research question, learners' and teachers' opinions were investigated. Learners' perceptions about listening difficulty and the use of videos in tests were obtained from the test-takers' questionnaire. The questionnaire was administered after each lecture in the ALC test. The data from the questionnaire was compared between the audio-only group and the video-based group of test-takers using factorial ANOVA analyses. To obtain teachers' perceptions, a questionnaire was developed. Teachers' perceptions were analyzed for their relationship to teachers' background (i.e., geographic region, education, and teaching-related experience).

Significance of the Study

The findings of the study informed the field of L2 listening assessment in both theoretical and practical terms. In theoretical terms, the study generated evidence for the

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

inclusion of video-based content-rich visuals into the L2 academic listening construct, begging refinements for existing construct definitions. In addition, the study offered practical insights into the use of innovative features in tests. As a unique contribution, the study introduced new approaches to (1) investigating effects of content-rich videos on academic listening comprehension, and (2) measuring the relationship between test items and the content of the videos, coined as video-dependence.

The study has been the first in the field to systematically investigate the role of videos that are rich in content-related graphical visual information in the L2 academic listening assessment construct. In this respect, the results of this dissertation study shed new light on whether the lack of content-rich videos in L2 academic listening tests should be regarded as construct underrepresentation and, thus, a threat to the validity of the listening scores. They also suggested possible improvements to the design of L2 academic listening tests. Finally, the study informed test developers and assessment specialists as to whether professional L2 teachers supported the innovation of using content-rich videos in L2 academic listening tests.

This evidence may lead the existing high- and medium-stakes L2 listening assessment practices to a new level, where resources afforded to test developers by video technology are successfully implemented, and assessment constructs are defined to be more representative of the visually rich L2 academic listening reality.

Role of the Researcher

The researcher served as a developer of the measurement instruments used in the study, a recruitment manager, and an online proctoring director. In addition, the researcher was a data coder and a data analyst for the study.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Assumptions

The following six assumptions were made for the study. First, the sample of ESL and EFL learners in the study was assumed to be similar in characteristics to the population of English language learners worldwide. Second, it was assumed that visual and auditory information is transmitted via two different sensory channels (i.e., the visual and auditory channels accordingly), but both types of information are processed simultaneously or near-simultaneously and are integrated during processing or after having been processed (Mayer, 2005). Third, it was assumed that the study's ALC test attained a high degree of authenticity through using lecture scripts and visual configurations from authentic lectures. It was also assumed that the delivery of lectures by actors did not significantly reduce the instrument's authenticity. Fourth, it was assumed that differences in scores between the audio-only and the video-based delivery modes could be attributed to the effects of videos despite possible use of test-taking strategies, such as elimination and guessing, by test-takers. Fifth, it was assumed that video effects on ESL/EFL listening comprehension were similar to effects on listening comprehension of any other second language. Sixth, it was assumed that all the participants had unimpaired listening abilities and vision.

Delimitations

This study was delimited in a number of ways. First, the sample of ESL and EFL learners was a convenience sample. This method did not give each individual in the population an equal chance to be chosen for the study, limiting the sample's representativeness of the target population. Second, L2 academic listening comprehension was assessed using multiple-choice comprehension questions only. L2

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

listening was measured by local and global questions. Third, only one academic text genre, lecture, was included to make inferences about test-takers' academic listening difficulty, leaving out listening texts belonging to other academic registers (e.g., an office hour conversation, a study group discussion). Only monologic non-interactive lectures were used. Finally, the study used quantitative research methods that generated product-oriented evidence for the inclusion of content-rich videos in the L2 academic listening construct (e.g., through the analysis of test-takers' scores). The study did not generate process-oriented evidence (e.g., through a qualitative analysis of test-takers' cognitive processes while taking the test). Such evidence could have offered deeper insights into the role of graphical visual information in academic listening.

Definition of Key Terms

Construct. An attribute, proficiency, ability, or skill that happens in the human brain and is defined by established theories (Brown, 2000).

L2 academic listening comprehension. Listeners' ability to process and understand incoming aural and visual input in their L2 (Rost, 2016).

Testlet. "A group of items related to a single content area that is developed by a unit and contains a fixed number of predetermined paths that an examinee may follow" (Wainer & Keily, 1987, p. 190). Reflecting this definition, a testlet in this study was a unit within a test that consisted of a listening passage followed by a number of comprehension questions.

Content-related visual cues. Video-based visual elements or sequences of elements, such as graphs, illustrations, and textual information, that illustrate, explain, or

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

describe the audio-based listening subject matter (Bejar, Douglas, Jamieson, Nissan and Turner, 2001).

Context-related visual cues. Video-based visual elements that are associated with verbal interaction unrelated to the listening content or the subject matter of a lecture, such as the setting and the speaker (Bejar et al., 2001).

Content-rich video. A digital recording of a lecture that has the following properties: (a) it sequentially displays several pictures and graphs, with each positioned side-by-side with the display of a lecturer, (b) the overall display time of the pictures, graphs, and the lecturer is about 20%, 40%, and 100% of the video length respectively, (c) the pictures and graphs are semantically congruous with the respective chunks of the auditory message, with the pictures fulfilling an illustrating function and the graphs serving as illustrators and/or organizers, occasionally providing some extra information (not assessed by the test), and (d) the pictures and graphs are intuitive and equally easy for viewers' interpretation.

Content-deficient video. Video that contains small or no amounts of content-related visual cues.

Viewing behavior. The extent to which test-takers orient to the video monitor during a video-based listening test (Wagner, 2007).

Organization of the Dissertation

This dissertation consists of five chapters. Chapter 1 has contextualized the main problem, stated the purpose of the study, given the overview of the methodology, the researcher's assumptions, the study's delimitations, definitions of key terminology, and the structure of the dissertation. Chapter 2 presents a critical synthesis of theoretical and

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

empirical literature pertaining to the topic of visually enhanced academic listening comprehension assessment, outlines research gaps, and states research questions. Chapter 3 describes the design and procedures of the study including data sources, data collection, and data analysis. Chapter 4 presents the findings for each research question in the study. Chapter 5 integrates the findings with the related literature and provides theoretical and practical implications for the field of L2 assessment.

Chapter 2

Literature Review

The rise of technology in the 21st century has had a profound impact on second language (L2) teaching. L2 academic listening is one area that has been particularly affected. One of the influences of modern technology is that L2 listeners can be provided with visual information that is normally present in academic listening target language use domains (TLU; Bachman & Palmer, 2010). This includes video-based L2 authentic or semi-authentic lectures such as graphs, PowerPoint presentations, blackboard notes, and illustrations in a textbook. In this sense, the teaching of L2 listening has caught up to the visually enhanced realities of the TLU domains (Lynch, 2011).

In contrast, the area of L2 academic listening assessment often fails to reflect these realities. Despite the growing ability to use video and multimedia technology in L2 assessments, standardized high-stakes L2 listening tests have operationalized academic listening almost exclusively as a visual-free skill (Kang, Gutierrez Arvizu, Chaipapae, and Lesnov, 2016). The only exception is using still pictures in tests such as the internet-based Test of English as a Foreign Language (TOEFL iBT). Pictures in these tests are not designed to aid in answering comprehension questions, and, thus, provide little assistance for test-takers. Recognizing the mismatch between the visually deficient acoustic input in tests and visually rich listening messages in authentic L2 classroom contexts, a growing number of researchers advocate for the inclusion of visuals in L2 listening tests in order for these tests to represent TLU domains more fully (e.g., Ockey, 2007; Suvorov, 2015a; Wagner, 2008).

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Visuals add information that requires processing on the part of L2 listeners (Mayer, 2005). This entails using the ability to decode visual information, which is called *visual literacy* (Debes, 1967). If a listening assessment included visual literacy, the construct could not be defined in the same way as it was for visual-free listening tests. The inclusion of visuals requires the redefining of the listening comprehension construct for L2 academic listening assessments.

The following section provides a brief introduction to the notion of a construct in L2 language assessment by describing how a construct is defined and how it is justified. These two themes of definition and justification guided the organization of the subsequent discussions in the literature review.

Defining and Justifying an Assessment Construct

Construct definition. A *construct*, also called a psychological construct, is “an attribute, proficiency, ability, or skill that happens in the human brain and is defined by established theories” (Brown, 2000, p. 9). In this broad sense, the word *construct* refers to any human behavior that cannot be directly observed and is used to study that behavior. In the assessment field, *construct* has a more specific meaning, namely a concept or characteristic that is measured by a test. An *assessment construct* is thus defined for purposes of measurement so that meaningful interpretations of test-takers’ scores can be obtained and used in a consistent manner (Chalhoub-Deville, 1997; 2003; Chapelle, 1998; 2011; Chapelle, Enright, Jamieson, 2008; 2010; Jamieson, 2014). Assessment-based interpretations are also prominent in Bachman and Palmer’s (2010) view of a construct as “the specific definition of an ability that provides the basis for a given assessment or assessment task for interpreting scores derived from this task” (p. 43).

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Therefore, both test development and test use are predicated on having a well-defined construct.

Defining a language assessment construct generally takes two steps. First, a construct needs to be specified based on the common theoretical understanding of a language ability or skill (Buck, 2001; Chapelle, 1998). The existing theory and research inform the componential structure of a construct such that the construct definition is meaningful for all stakeholders. For this, a contemporary theoretical model of a particular listening ability is often used as a frame of reference. Second, the definition is refined by the knowledge of TLU domains and situations (Buck, 2001). This ensures that the conceptualization of a language skill mirrors the practices and forms of language use in TLU situations. For example, a construct of L2 academic listening at universities would have to include the ability to process informationally dense stimuli, such as lectures. The degree of correspondence between a test construct and test tasks to the TLU situations determines test authenticity (Bachman & Palmer, 1996, p. 24).

Construct justification. To be accountable to stakeholders, test developers need to justify the use of a particular assessment. This is done by providing evidence in support of the interpretations and decisions made on the basis of the assessment. Specifically, it requires test developers to prove that their instrument measures the construct that it claims to measure, the concept broadly referred to as *validity* (Cronbach, 1971). This purely construct-oriented validity model was further amplified with the dimension of social consequences and implications (Messick, 1989). Messick defined validity as “an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of inference and

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

actions based on test scores or other modes of assessment” (p. 13). To ensure this appropriateness, a measure (i.e., an operationalization of the construct) should reliably generate scores that are indicative of the degree to which the measured construct is developed in an individual.

The contemporary view of validity is largely based on the Kane’s (2004) argument-based framework. According to Kane, validation requires the building of a twofold argument consisting of an interpretive argument and a validity argument. The interpretive argument articulates the intended interpretations of test scores, or inferences. The validity argument presents evidence for the accuracy of these inferences. For example, if an inference posits that test scores are consistent across test-takers’ groups (interpretive argument), a reliability analysis would be one way to provide evidence for this inference (validity argument). As informed by the argumentation model proposed by Toulmin (2003), each inference is associated with the aggregation of *warrants*, *backings*, and *rebuttals*. Warrants are assumptions showing how the data support the *claim* of the inference. Referring to the above-mentioned consistency example, the warrant would be that observed scores are consistently awarded over relevant parallel versions of the test. Evidence, or backing, for this warrant can be the reliability analysis results. A rebuttal, on the other hand, is a condition that challenges the warrant. For the discussed example, an implied rebuttal would be that the observed scores are not internally consistent. In the presence of a strong backing for the warrant, rebuttals are refuted.

Kane’s validity framework has evolved to have five inferences, including *evaluation*, *generalization*, *explanation*, *extrapolation*, and *utilization* (Bachman, 2005; Bachman & Palmer, 2010; Chapelle, 2012; Chapelle et al., 2008; Kane, 2004).

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Recognizing the importance of the language use domain specification, Chapelle et al. (2008) added one more inference, *domain definition*, to this chain of inferences. They used this expanded framework to construct a validity argument for the Test of English as a Foreign Language (TOEFL iBT). Each inference in Kane's framework moves the argument from data to a claim, reflecting Toulmin's argumentation theory. The claim of a preceding inference becomes the data for a subsequent inference. As mentioned above, inferences contain corresponding warrants and assumptions, which are legitimized by backings. Starting with domain definition, information about each inference in a language test validation argument is summarized in Table 2.1 below with regards to respective data, claims, warrants, assumptions, and backings. The summaries in Table 2.1 were primarily informed by Aryadoust (2013), Bachman and Palmer (2010), Chapelle et al. (2008), Gruba (2014), and *the Standards for Educational and Psychological Testing* (1999/2014; henceforth, *the Standards*).

As reflected in Table 2.1, the domain definition inference links the TLU domain characteristics to test content and tasks. It warrants that test behavior elicited by test tasks is reflective of test-takers' skills typically used in assessment situations in the authentic contexts. This determines test authenticity, which has been viewed as a highly desirable quality of a test (Bachman & Palmer, 1996; Cumming & Maxwell, 1999; Douglas, 1997). The assumptions for the test domain inference are backed by evidence from domain analyses, which informs test developers of typical language abilities required in the TLU domain and typical assessment tasks (Mislevy, Steinberg, & Almond, 2003).

Table 2.1

Inferences in Kane’s Interpretive and Validity Arguments for an L2 Test

Inference	Data	Claim	Warrant	Assumptions	Backings
Domain definition	Target language use domain (TLU)	Test content and tasks are representative of the TLU.	Test-takers’ performance reflects their skills in contexts representative of the TLU.	<ul style="list-style-type: none"> a) Language skills and processes needed for the TLU are identified. b) Assessment tasks typical of the TLU are identified. c) Test tasks can be created to reflect (a) and (b). 	<ul style="list-style-type: none"> a) Domain analysis: theory-driven identification of the required language abilities b) Domain analysis: expert-based identification of typical TLU assessment tasks c) Development of test content and tasks informed by (a) and (b)
Evaluation	Test content and tasks	Observed scores reflect the relevant aspects of performance.	Observed scores reflective of the TLU abilities are consistently awarded.	<ul style="list-style-type: none"> a) Appropriate scoring rubrics and/or methods are used. b) Test performance is not affected by administration conditions. c) Items are psychometrically appropriate for norm-referenced decisions. 	<ul style="list-style-type: none"> a) Clear and reliable rubrics and/or scoring methods b) Properly controlled testing conditions c) Item analysis
Generalization	Observed scores	Observed scores predict expected scores across parallel tasks and forms.	Test scores are generalizable to expected scores in the target test domain.	<ul style="list-style-type: none"> a) The number of items is sufficient for stable estimates of test-takers’ performance. b) The configuration of tasks is appropriate for the intended interpretation. c) Test administration can be easily replicated for other samples. d) Proper equating and scaling methods are used. 	<ul style="list-style-type: none"> a) Reliability and generalizability studies b) Reliability analyses for different task configurations c) Evidence based on test specifications d) Equating/scaling analysis

(continued)

Table 2.1 (continued)

Inference	Data	Claim	Warrant	Assumptions	Backings
Explanation	Expected scores	Theoretical construct has been properly defined.	Observed scores are attributed to the defined construct.	<ul style="list-style-type: none"> a) The linguistics knowledge and processes vary in keeping with theoretical expectations. b) Test performance relates to behavior on other measures in keeping with theoretical expectations. c) Test internal structure is consistent with the theory. d) Item difficulty is affected by construct-relevant factors (e.g., item type). e) Test performance is not affected by construct-irrelevant variables (e.g., gender). f) The construct definition is supported by test stakeholders. 	<ul style="list-style-type: none"> a) Cognitive processing analysis b) Convergent or divergent correlational studies c) Correlational and factor analysis studies d) Analysis of factors affecting item difficulty; convergent and discriminant evidence e) Comparison of group differences (no differences expected) f) Studies into stakeholders' perceptions justifying the developed construct definition and its operationalization
Extrapolation	Construct	Target score represents performance in the authentic context.	Test performance accurately predicts behavior in the TLU.	<ul style="list-style-type: none"> a) There is a relationship between test performance and behavior on the criterion in authentic contexts. b) Existing research on relations of similar instruments to the criterions is analyzed. 	<ul style="list-style-type: none"> a) Analyses of concurrent and/or predictive test-criterion relationships b) Meta-analyses of previous test-criterion studies
Utilization	Target score	Target score reflects test-takers' ability to use the language in authentic contexts and is useful for proper decision-making.	Decisions made based on the target score are valid.	<ul style="list-style-type: none"> a) The meaning of the score is clearly interpretable by teachers, admission officers, and test-takers. b) The test has a positive washback effect on language teaching and learning. c) The test does not have unintended consequences. 	<ul style="list-style-type: none"> a) Available materials guiding score interpretations and use b) Washback studies c) Logical or empirical studies of unintended consequences

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

The evaluation, or scoring, inference takes us from test content to observed test scores. It warrants that test-takers' performance is consistently observed and scored. The assumptions for the evaluation inference are backed by using scoring methods informed by existing theory and assessed by researchers. Evidence from prototyping and pilot studies are also needed to set and control testing conditions, such as time constraints, interface, and instructions. Finally, psychometric qualities of test items need to show that scores are appropriate for norm-referenced decision making. In norm-referenced situations, test items range in difficulty and distinguish among several ability levels of test-takers. To show this, item difficulties and discriminations are inspected using either classical or item response theory analyses.

The generalization inference links the observed score to the expected score, which is a hypothesized score that a test-taker would be expected to get in a similar testing situation. It warrants that observed scores can be reproduced across multiple test administrations, forms, and similar conditions. The assumptions for the generalization inference are backed by reliability analyses, evidence for test administration replicability, and equating/scaling analyses. Besides calculating an overall internal consistency index, reliability may be examined in relation to different subsets of test items, in a quest to determine the most reliable configuration of a test. The replicability of test administration is ensured by having thorough test specifications and item development guidelines, or by creating arsenals of trialed tasks to be used in parallel test versions. The equating and scaling analyses seek to eliminate the effects of unintended differences in test form difficulties on the expected score (Dorans, Moses, & Eignor, 2010).

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

The explanation inference links the expected score to the construct. Its warrant is that the expected score measures the defined construct. The assumptions for the explanation inference are backed by cognitive processing, correlational, item difficulty, group differences, and stakeholders' perceptions studies. Qualitative cognitive processing studies help to confirm that test-takers indeed use the skills and processes theorized as parts of the construct definition. Correlational studies quantitatively analyze convergent relationships with measurements of similar constructs as well as divergent relationships with measurements of different constructs. Item difficulty studies seek to confirm the theoretical expectations of differences in item difficulties due to construct-relevant factors, such as item type or measured subskill. Group differences studies often investigate the effects of content-irrelevant factors, such as demographic factors. The absence of group differences lends support for the test construct. As pointed by Gruba (2014), stakeholders' perceptions on the role of relevant factors in assessment constructs serve as another piece of evidence for developing and justifying construct definitions.

The extrapolation inference links test-takers' behavior on the test construct to their behavior in projected authentic contexts in the TLU domain. It warrants that the test target score accurately predicts test-takers' performance in the corresponding TLU domain. The assumptions for the extrapolation inference are backed by evidence of test-criterion relationships. A criterion is "a measure of some attribute or outcome that is operationally distinct from the test" (*the Standards*, p. 17), but reflective of intended test uses. For instance, an academic language proficiency test may have test-takers' first-semester GPAs as one possible criterion. Test performance is expected to predict criterion performance, indicating the test's capacity to cause appropriate interpretations,

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

uses, and consequences. Another indication of test score generalizability to authentic situations could be provided by a meta-analysis of test-criterion relationships studies that used similar tests in relevant local contexts (*the Standards*, 2014).

The utilization, or decision, inference links the target score to its uses. It warrants that the test score lead to valid decision making for test-takers. The assumptions for the utilization inference are backed by comprehensive documentation on the test score interpretation and use as well as by evidence from washback studies. The former ensures that the target score is properly and consistently interpreted by stakeholders while the latter is expected to show the test's positive impact on learning and teaching. Logical or empirical evidence for absence of unintended negative test consequences could serve as an additional backing.

Kane's argument-based approach has two clear advantages over the other approaches to validity. First, it has a capacity to integrate both previous and contemporary ideas about validity of test score interpretation and use in one logical comprehensive framework, as opposed to mere evidence-gathering. This framework accounts for the retired but valuable concepts of content, criterion, and construct validity, as well as for the totality of the validity evidence types explicated in *the Standards*. Yet unlike the previous frameworks and *the Standards*, the argument-based approach avoids overreliance on construct validity. Instead of overemphasizing construct validity, an interpretive argument outlines the inferences, warrants, and assumptions that motivate score interpretation and use. It helps to remove "the enormous burden that might otherwise be placed on an imprecise theoretical construct" (Chapelle et al., 2010, p. 12). Second, the argument-based structure gives advantages to Kane's validity framework. A

well-developed interpretive argument is essentially a set of instructions guiding the evidence-gathering process. “The interpretive argument also provides a basis for identifying the most serious challenges to a proposed interpretation – challenges that expose weaknesses (e.g., hidden assumptions) in the interpretive argument” (Kane, 1992, p. 9). Due to these and other advantages, Kane’s validity framework has received recognition and is now widely used in the L2 language assessment field (Chapelle & Voss, 2014). These strengths also motivated the use of Kane’s framework in this dissertation study.

Summary. Out of the six types of validity inferences, test domain is most closely associated with properly defining an assessment construct and developing authentic test content and tasks. The test domain inference uses domain analysis to fuel the two steps in the process of developing a construct definition discussed above. First, it ensures that the test construct is informed by the contemporary theory about a language skill. This is done by conducting literature reviews of existing expert-informed theoretical models and conceptualizations of the language skills. Second, it evaluates if the construct mirrors the corresponding authentic language uses in the TLU domain. This evaluation usually includes surveys of language tasks in the TLU domain.

The explanation inference is most closely associated with justifying an assessment construct (Aryadoust, 2013). It claims that “the test indeed measures what its underlying theoretical construct claims to measure” by providing evidence that the construct is sufficiently represented, and that construct-irrelevant variance is minimized (p. 31). The backing for the explanation inference often includes cognitive analyses, item difficulty analyses, correlational studies, factor analyses, and group differences studies (Chapelle et

al., 2008; Gruba, 2014). One way to justify a construct would be to compare item difficulties under two conditions – when the construct is underrepresented versus when the construct is sufficiently represented (according to the theory), anticipating significantly different performances under these conditions. For instance, if theory holds visual information as part of listening and advocates for its positive influence on comprehension, a visual-free listening assessment construct would be underrepresented. This underrepresentation may empirically manifest itself in higher item difficulties, compared to items difficulties generated under a visual-inclusive listening construct. Gruba (2014) also argued that stakeholders' perceptions studies could help to obtain evidence supporting the construct-related theory. L2 learners or teachers' perceptions about the nature of a particular language skill should be taken into account for developing the assessment construct (Bachman & Palmer, 2010; Gruba, 2014).

Organization of Review. As part of developing an interpretive argument for the inclusion of content-rich visuals in the L2 academic listening assessment construct, four steps were taken. First, the theories of the L2 listening skill needed to be reviewed and synthesized, which generated a preliminary definition of the L2 academic listening construct. Second, the domain of the L2 academic listening needed to be described. This further refined the construct definition regarding the role of visual information. The first and the second steps provided backings for the test domain inference in the interpretive argument. Third, studies comparing item difficulty (or test performance) under a visual-free versus a visual-inclusive condition needed to be reviewed. This review provided empirical evidence as to whether visual information constitutes construct-relevant variance. Fourth, research into stakeholders' perceptions about the role of visual

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

information in L2 listening had to be reviewed. It showed whether stakeholders approved of including visual processing components into the L2 academic listening construct. The third and the fourth steps provided backings for the explanation inference in the interpretive argument. Each of these steps is addressed in the following four sections of the literature review. A comprehensive summary of research gaps and research questions conclude this chapter. The structure of the remainder of the literature review is depicted in Figure 2.1 below.

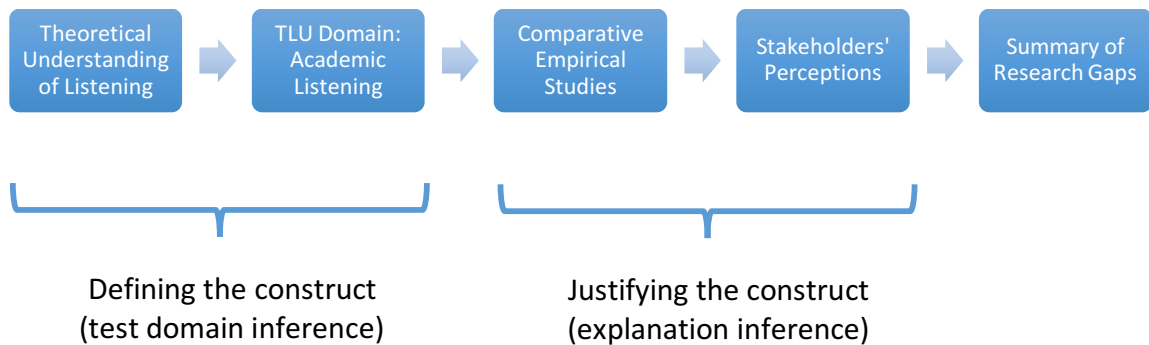


Figure 2.1. Organization of the literature review.

Theoretical Understanding of L2 Listening Skill

In this section, the first step in ratifying the visual-inclusive L2 academic listening assessment construct is addressed. The section elaborates on current scholarly views of academic listening, which informed the construct definition. A theoretical model of L2 listening is discussed first. The role of visual information in L2 listening is addressed next. The section then reports on the evolution of L2 listening definitions during the last 80 years and concludes with a brief summary.

Theoretical model of L2 listening. Theoretical models of the L2 listening process derive from the existing theoretical understanding of first language (L1) listening. Unlike L1 listening, L2 listening can be constrained by listeners' language proficiency

and lower flexibility of psychomotor skills. After the critical period in L2 language acquisition, human brain often lacks flexibility for native-like performance in productive and perceptive skills, including listening. However, it is generally accepted that L2 listening is not fundamentally different from L1 listening in terms of cognitive processes involved (Buck, 2001; Rost, 2005). As a result, the current conceptualizations of the L2 listening process largely mirror L1 listening research and theory (e.g., Anderson, 2000; Clark & Clark, 1977; Cutler & Clifton, 1999; Kintsch, 1998).

There is a relatively small number of fully elaborated cognitive models of L2 listening (i.e., Bejar, Douglas, Jamieson, Nissan, & Turner, 2001; Field, 2013; Rost, 2016; Vandergrift & Goh, 2012). To varying degrees, they all reflect the frequently cited listening framework of Anderson (2000). Anderson described listening as a three-operation process that includes *decoding* (recognition of sound signals), *parsing* (syntactic segmentation of an utterance), and *utilization* (building a mental representation using the existing knowledge). In addition to these three fundamental operations, the current models also include listener response and strategy use dimensions (Bejar et al., 2001; Rost, 2016; Vandergrift & Goh, 2012), affective and social dimensions (Rost, 2016), and more fine-grained views of utilization (Field, 2013; Vandergrift & Goh, 2012). Field's model is particularly useful because it (a) has a balance between perception and understanding, (b) takes fuller and more logical account of higher-level processes, and (c) excludes components that are less relevant to university lecture comprehension (e.g., socio-affective dimensions).

Partitioning Anderson's (2000) original *utilization* listening component, Field (2013) posited that L2 listening is a five-operation process. These operations include

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

input decoding, lexical search, syntactic parsing, meaning construction, and meaning representation. According to Field, the first three operations constitute lower-level processes while the latter two are considered higher-level processes. Each of these operations, as described by Field, is explained below.

Input decoding. The process of transforming an acoustic signal into a set of abstract representations is called input decoding. Abstract representations include syllables or groups of syllables that correspond to the phonological system of the target language, helping to make sense of what is heard. This process requires decoding at the phoneme and syllable levels. Also, it largely depends on listeners' ability to recognize lexical stress. Focal stress may signal clearer boundaries, leading to a more meaningful transformation of a string of phonemes into a set of syllables. Thus, the product of input decoding is a string of stress-marked syllables and is clearly a function of the listener's phonological knowledge of the target language.

Lexical search. From the newly generated phonological string, the L2 listener identifies lexical items that best correspond to the spoken forms of learned words, which are stored in long-term memory. The listener constantly maps segments of the string of syllables to a number of likely word-level candidates. As more phonological cues become available, the number of these candidates reduces gradually until the best match is found. This process, called lexical search, is cued by lexical stress recognition and mental lexicon capacity including but not limited to the awareness of word frequency, collocations, and meaning. The product of lexical search is a string of content and function words, with meaning allocated to some content words and yet undecided for others. An outcome of lexical search is determined by the listener's lexical knowledge.

Syntactic parsing. Information in the lexical string needs to be analyzed against other elements of text in which it occurs so that a grammatical structure is imposed on the word group. This process happens online while phonological and lexical information continues to be received and is referred to as syntactic parsing. In addition to mapping the lexical string onto a grammatical model, syntactic parsing narrows down the range of possible meanings for certain words and helps to predict future word-level or phrase-level structures. Syntactic parsing is a function of the listener's syntactic knowledge. It can also be assisted by recognizing intonation contours of phrases, which mark prominent grammatical elements and speakers' intentions. The end product of the process is an informed decision on the phrase meaning and building a proposition, or an abstract idea, of the incoming message. The proposition is continuously updated until all language chunks have been parsed. Eventually, the proposition is stored in a non-linguistic form in short-term memory and is not yet integrated with the context of listening (Lyons, 1977). In other words, the proposition at this stage is a raw meaning of the incoming message.

Meaning construction. To grasp the full significance of the message, this bare meaning needs to be linked to the context in which the act of listening occurs. Meaning construction supplies context-specific information to the listener and activates context-related schema. This process usually feeds the listener with four types of information including pragmatic, contextual, semantic, and inferential information. Using pragmatic knowledge, the listener interprets the speaker's purposes and intentions. Next, the listener enriches the proposition with the contextual information using (a) world knowledge, knowledge of the speaker, and knowledge of the situation, and (b) recall of previous segments of the listening discourse. Employing knowledge of the ideas in the listening

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

message (i.e., schema; Barlett, 1932), the listener is able to better interpret those ideas. This knowledge also helps to make inferences about the implied details not mentioned by the speaker. The final product of meaning construction is a mental model enriched by the listener's pragmatic understanding, world knowledge, and previous discourse recall related to the context of listening.

Discourse construction. During discourse construction, the listener analyzes the relevance of new information and makes judgements about its consistency with the previous information from the listening message. Drawing upon the work in reading comprehension processes, Field (2013) suggested that discourse construction consists of four sub-processes including selection, integration, self-monitoring, and structure building. Selection helps the listener to take notice of the information central to the topic or the speaker's goals. Self-monitoring enables the listener to evaluate this information for consistency with the previous discourse units. If consistency is confirmed, the information is linked to the previous discourse and added to the mental model as part of the integration sub-operation. In the structure building sub-operation, the listener constructs the hierarchical pattern of the discourse by identifying major and minor points of the listening content and storing this structure in the short-term memory. Discourse construction utilizes the listener's world knowledge, previous discourse recall, and familiarity with various discourse types. As a result, the listener's is equipped with a proper discourse representation that complements the mental model.

Summary of the model. Field's model is in line with other influential views of L2 listening. It does not contradict the conventional view of listening as a combination of decoding and comprehension (Wolff, 1987). Similarly, it agrees with the popular

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

distinction between bottom-up and top-down processes of listening (Morley, 2001; Morley, 2007; Vandergrift & Goh, 2012). Bottom-up processes are reflected by Field's speech perception, lexical search, and syntactic parsing. Top-down processes are believed to involve interpreting the decoded message through the filter of contextual and prior knowledge. Thus, they are parallel to Field's higher-level processes of meaning construction and discourse construction.

Two characteristics of Field's model should be noted. First, the five operations in the model are not necessarily sequential. Even though higher-level processes may be dependent on the successful production of the lower-level processes, the sub-processes within each of these levels likely happen simultaneously. Moreover, L2 listeners can be employing lower-level processes for a newly coming chunk of input while still working on the meaning and discourse construction with regards to the previous chunk. Thus, the processes operate "in close conjunction" (Field, 2013, p. 101) and in a recursive fashion. Second, the model is a singular-mode prototype of L2 listening. It does not explicitly allocate a role to visual information in the listening process. That said, the model may be useful for estimating this role with respect each of the model components.

Visual information in L2 listening. To estimate the role of visuals in L2 listening thoroughly, it is necessary to take account of different types of visual information. This would lead to a more meaningful analysis of visuals' impact on both lower-level and higher-level processes in L2 listening.

Types of visual information. Visual information usually refers to an object or a group of objects that a person can observe using their sense of sight. The meaning of *a visual* is normally more specific and refers to "a picture, piece of film, or display used to

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

illustrate or accompany something” (“Visual,” 2017), often for instructional purposes.

Visuals can range from being a still picture, through a moving picture, to a video (Ockey, 2007; Bejar et al., 2000). Videos are a kind of media that combine visual and audio elements in close temporal sequence (Wetzel, Radtke, & Stern, 1994).

A number of classifications of visual information can be found in the literature. Four are particularly relevant to L2 assessment. The classification that is most commonly used in the L2 listening assessment field is based on the distinction between context and content visuals, first introduced and applied by Bejar et al. (2000) and further developed by Ginther (2002). According to Bejar et al., context visuals contain information associated with verbal interaction unrelated to the listening content (i.e., visuals showing features of listening “situation” including the setting and the speakers). Context visuals are further subdivided into visuals about (a) setting, such as a visual showing a classroom; (b) participants such, as a visual showing a lecturing professor; and (c) text type, such as a visual showing a dialogue.

In contrast, content visuals provide important information on the actual subject of the auditory stimulus. Bejar et al. (2000) classified content visuals based on their relationship to the listening input dividing them into four subtypes including visuals (a) replicating the oral stimulus, such as a sentence that was verbalized in the auditory stimulus; (b) illustrating the oral stimulus, such as a picture of a dinosaur described in the stimulus; (c) organizing information in the stimulus, such as a diagram of the discussed process; and (d) supplementing the oral stimulus, such as a visual presenting new information that does not have a match in the listening message.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Another classification, introduced by Rost (2016), was based on the difference between nonverbal cues (e.g., gestures, movements, facial expressions) and visual aids (e.g., related drawings, schemes, or presentation slides). Accordingly, Rost distinguished between two main types of visuals, namely kinesic and exophoric. Kinesic visuals define nonverbal cues and include “baton signals” (i.e., hand and head movements for emphatic purposes), directional gaze (i.e., an eye movement pointing to a particular object or part of discourse), and guide signals (i.e., systematic idiosyncratic gestures usually with no particular meaning). In contrast, exophoric visuals are references for the spoken text and are essential for understanding informationally heavy listening messages such as academic lectures (e.g., a drawing or text on the blackboard).

Kinesic versus exophoric classification is somewhat similar to context versus content classification described above. Kinesic visual information can be seen as a sub-category within context visuals, covering the situational cues associated with the speaker’s body language but not necessarily with the setting itself. The category of exophoric visuals largely overlaps with the content type.

Walma van der Molen (2001) based her taxonomy of video-based visuals of news reports on the degree of semantic overlap between the audio and video channels. The four categories of this taxonomy are as follows: Direct, indirect, divergent, and talking head. If an audio and an accompanying video contained semantically redundant information, they fell in the direct category. The indirect audio and video content would be related only partly. Audio and video were of the divergent type if their contents were not related or contradictory. The fourth category, talking head, included videos displaying the upper part of a reporter’s body only.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

A classification of visuals associated with the register of conference presentations was generated by Rowley-Jolivet (2002). The four types included *graphical* (e.g., diagrams, schemes), *figurative* (e.g., photographs), *scriptural* (i.e., based on text), and *numerical* (i.e., based on numbers). Figurative and scriptural visuals were further categorized as polysemic, or allowing for several possible interpretations, while graphical and numeric visuals were designated monosemic, or unambiguous.

The existing classifications of visual clues have some major limitations. It seems that most of these classifications discriminate between non-verbal cues (partly or fully corresponding to context visuals, kinesic visuals, and talking head in the classifications above) and content-related cues (partly or fully corresponding to content, exophoric, and direct categories of visuals in the classifications above). However, these two categories are not mutually exclusive. Attempts to categorize more complex visuals, such as videos, would often be unsuccessful because such visuals normally contain elements belonging to different categories. For example, a lecture video would combine the elements of both context and content, kinesic and exophoric, direct and indirect, or graphical and scriptural. As a result, the investigation of video effects based on these classifications would be meaningful only in a rare scenario of using videos with no mixture of visual types. In addition, the existing visual categories fail to reflect the degree of visuals' contribution to the understanding of the listening message. It may be more meaningful to place visuals on a continuum representing the degree to which their presence is helpful for comprehending the listening stimulus.

Theorized benefits of visual information. Existing models related to multimodal information processing shed light on the role of visual information in listening. One such

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

model stems from the dual coding theory, which postulates that learners process incoming stimuli by means of (a) decoding words and (b) decoding images (Paivio, 1979; 1991). Accordingly, the decoding of incoming information is mediated by the two distinct channels, namely the visual and auditory channels. Visual codes form mental representations of images, which usually closely mirror physical characteristics of observed objects (Sternberg, 2003). Verbal codes mentally represent words that are heard or read by a person. The two systems are believed to operate simultaneously and reinforce each other, generating organized knowledge units and facilitating language learning. The theory has been supported by empirical evidence showing that the presence of visuals increased the memory of verbal input and enhanced learners' reading and listening comprehension, and writing skills (e.g., Brunye, Taylor, & Rapp, 2008; Paivio, 2006; Paivio & Lambert, 1981; Purnell & Solman, 1991; Yang, 2014).

The visual/verbal distinction was also fundamental for the theory of generative multimedia learning developed by Mayer (2002; 2005; 2009). It posits that each of the two processing channels (i.e., visual and auditory) has limited capacity and is most active when a learner is focused on the incoming auditory-visual signals. The information from both channels is stored in auditory working memory and visual working memory accordingly. This information is then used to construct two corresponding mental models of the auditory-visual message – verbal mental model and pictorial mental model, which are eventually integrated. Similar to Paivio's dual coding theory, Mayer's theory maintains that visual information, if congruent with an auditory input, is an additional resource helping to decode and organize the latter (Vandergrift & Goh, 2012).

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

According to the connectionist cognitive processing model introduced by Kintsch (1998) and further revised by Bejar et al. (2000), the listener receives and processes both aural and visual input simultaneously, constantly modifying the interpretation as the input goes on. During this stage, called listening, a set of beliefs, or propositions, about the auditory and visual stimuli are cognitively constructed with the help of the listener's situational, linguistic, and background knowledge. The propositions are then used to form a response to the stimulus, depending on the situation (e.g., the listener's answer to a comprehension question). This model reflects complexities of the listening process including its many-faceted structure and its interactional and visual-inclusive character.

It is also useful to estimate the functions of visual information in specific L2 listening operations, or sub-processes, in Field's (2013) model. The nature of these functions is likely to depend on the type of visuals viewed by a listener, namely whether visuals are more context-based or content-oriented. Context visuals are generally considered to be helpful for listening comprehension. Specifically, they are thought to facilitate speech perception through lip reading (Green, 1998; Massaro, 1987), confirm or disconfirm the linguistic meaning (Rost, 2016), situate the auditory stimuli within the context of a given listening message (Bejar et al., 2000; Field, 2008; Rubin, 1995), activate listener's background knowledge and schema (Rubin, 1995), and reduce listening effort (Picou, Ricketts, & Hornsby, 2011). However, Bejar et al. and Rost warn that gestures may also be detrimental for comprehension if they are inconsistent with the given linguistic input or listener's cultural expectations.

It seems that visual information that is more context-oriented is a facilitator of both the lower- and higher-level processes in Field's L2 listening model. It affords the

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

listener a possibility of lip reading and interpreting articulatory movements, which may assist input decoding at the phoneme and syllable levels as well as lexical search. This, in turn, may bear a constructive impact on syntactic parsing. In addition, context-oriented visual information, including the speaker's non-verbal cues and situational cues, aid in interpreting the speaker's intentions and activating the listener's world knowledge. These are the core sub-processes at the higher-level meaning construction listening stage. It should be noted, however, that beginner L2 listeners may fail to take advantage of context-related visual benefits due to their limited capabilities of attending to linguistic and visual sources simultaneously (Gruba, 2004).

Similarly, content-oriented visuals make listening comprehension easier because of the potential to replicate, illustrate, and visually organize the auditory stimulus (Bejar et al., 2000). Rost (2016) emphasized the critical role of exophoric visuals for comprehending heavily loaded listening messages because they provide a reference for the spoken text, and, thus, are essential for successful comprehension. Field (2008) also argued that such visuals help listeners to comprehend spatial relationships between the speaker and the listener. According to Rowley-Jolivet (2002), scriptural (textual) visual elements can reduce memory load and difficulties stemming from listeners' limited L2 abilities and increase the global comprehension of the message.

Common sense dictates that different content visuals will exert positive influences on different listening sub-processes. For example, while photographs or video scenes in a news report could in some way promote speech perception, they would probably be most instrumental for meaning construction. Alternatively, text-based PowerPoint slides during a lecture will likely visualize spoken phonemes, lexical items, and syntactic structures in

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

a text form. This will tremendously assist the listener with input decoding, lexical search, and parsing. PowerPoint-based visuals may also delineate the structure of a talk, promoting discourse construction. In other words, the impact of content visuals on L2 sub-processes may depend on the form of a visual used by the speaker. It is hypothesized that this impact will make the L2 listening process easier; however, this hypothesis may not hold true for lower-level L2 listeners.

Factors affecting the role of visual information. Proficiency. As alluded to above, L2 learners' proficiency is one factor affecting the function of visual information in listening comprehension. For L2 listeners with limited linguistic, pragmatic, and cultural knowledge of the target language, the simultaneous integration of listening processes and visual input may be difficult. The lack of resources may offset the advantages of the visual input and turn visual interpretation into a burdensome task (Gruba, 2004). Sometimes L2 listeners need to use other language skills, such as reading, to interpret visual input (e.g., text slides). If listeners have low reading skills, such visuals may end up being useless or distracting.

Visual literacy. Visual literacy is another factor that shapes the role of visual information in the L2 listening process. The concept of visual literacy was developed by Debes (1967). Debes defined it as a person's ability to "translate from visual language to verbal language and vice versa" (p. 27). The modern view of visual literacy includes the abilities associated with visual reasoning, visual thinking, visuals association, visual meaning construction, and using visual conventions among others (Avgerinou & Pettersson, 2007). Despite living in a digitally-rich era, many individuals lack these abilities. For example, Malamitsa, Kokkotas, & Kasoutas (2008) found that interpreting a

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

graph was not intuitive for Greek college students. Similarly, Beaudoin (2016) reported on college students' inability to interpret historical pictures. This leads scholars to recognize the need of visual literacy instruction at different educational levels. For example, the National Council of Teachers of English (NCTE) in the USA argues that modern students should have the ability to analyze and evaluate "multimedia texts" ("NCTE Framework," 2013). The Association of College and Research Libraries (ACRL) advanced this argument by publishing the Visual Literacy Competency Standards for Higher Education ("ACRL Standards," 2011). These standards stipulate high expectations from American college students, including the ability to identify visual information and interpret its meaning (Hattwig, Bussert, Medaille, Burgess, 2012). Visual literacy is also becoming a part of educational curricula in other countries, which raises hopes that individuals' variability in visual decoding is being globally reduced (e.g., Wagner & Schönau, 2016).

Viewing behavior. The effect of visuals on the L2 listening process may also depend on how attentively listeners view visual information at hand, the notion that is referred to as *viewing behavior* (Ockey, 2007; Suvorov, 2015a; Wagner, 2007, 2010a). Viewing behavior can range from completely ignoring visual input to fully immersing into it (Cubilo & Winke, 2013; Wagner, 2007) and depends on individual preferences. For instance, Wagner's (2007) study found that test-takers spent from 37 to 90% of the listening time oriented to videos, and that watching rate affected learners' listening comprehension. Viewing behavior, in turn, may be a function of listening proficiency, the degree of visual literacy, learners' motivation or preferences. It is, thus, reasonable to expect test-takers' listening processes to vary in response to their viewing behavior.

Evolution of L2 listening definitions. Recognizing the support that relevant theory lends to the role of visual information as part of the L2 listening process, it is informative to analyze how the L2 listening skill has been defined in relation to visual information of different types. Table 2.2 provides a historical account of existing listening definitions proposed by theoreticians from the 1920s to 2016. The table shows scholars' names and years of their work, definitions of the listening skill, and roles allocated to visual information in the definitions.

Looking at the table, we can see that there is one commonality among the definitions. All the scholars seem to agree that constructing meaning from an incoming acoustic verbal signal is a core process in the L2 listening. This singular process largely sufficed in the early definitions (e.g., Chastain, 1978; Fries, 1947; Furness, 1952; Lado, 1961; Rankin, 1928). More recent scholars elaborated on the cognitive processes of L2 listening. As a result, many other dimensions were added to L2 listening definitions including activating schema, interpreting the speaker's intentions, integrating linguistic, pragmatic, and semantic processes, having the knowledge of discourse, using metacognitive strategies, and giving a listener's response (e.g., Buck, 2001; Field, 2008; Richards, 1983; Rubin, 1995; Ur, 1984). In this sense, the modern definitions largely reflect Field's (2013) theoretical model of the L2 listening skill, which was reviewed on pages 8-13. The model acknowledges the functions of the linguistic, pragmatic, semantic, and discourse processing as well as of schema in one or more of the five operations (i.e., decoding, lexical search, syntactic parsing, meaning construction, and discourse construction).

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Table 2.2

Historical Overview of L2 Listening Definitions

Author	Definition	Role of Visuals
Rankin (1928)	The ability to understand spoken language (p. 623)	No role allocated to visual clues
Fries (1947)	The understanding of the stream of speech requiring the mastering of both sound segments and covering intonation patterns (p. 24)	No role allocated to visual clues
Furness (1952)	The ability to understand a foreign language when it is spoken (p. 124)	No role allocated to visual clues
Lewis (1958)	The process of hearing, identifying, understanding, and interpreting spoken language (p. 89)	No role allocated to visual clues
Lado (1961)	Recognition control of the signaling elements of the language in communication situation (p. 206)	No role allocated to visual clues
Chastain (1976)	The ability to discriminate between the significant sound and intonation patterns of the language, perceive an oral message, keep the communication in mind while it is being processed, and understand the contained message (pp. 287-293)	No role allocated to visual clues
Richards (1983)	Three related levels of discourse processing appear to be involved in listening: propositional identification, interpretation of illocutionary force, and activation of real world knowledge (p. 220)	Facial, kinesic, body language, and other non-verbal cues to decipher meaning
Ur (1984)	The process of attending to and understanding the incoming message through both the aural and visual channel (pp. 2-20)	Non-verbal signals and environmental cues are part of L2 listening
ILA (An ILA Definition, 1995)	The active process of receiving, constructing meaning from, and responding to spoken and/or nonverbal messages (p. 1)	Non-verbal cues facilitate listening comprehension
Rubin (1995)	An active process in which listeners select and interpret information which comes from auditory and visual cues in order to define what is going on and what the speakers are trying to express (p. 7)	Interpreting visual cues is an essential part of the listening process
Buck (2001)	An active process of constructing meaning from the incoming sound (p. 31)	Non-verbal cues have the potential to influence listener's interpretation of the listening message (p. 48).
Flowerdew & Miller (2005)	Processing phonological, syntactic, semantic, pragmatic, and kinesic information (p. 45)	Kinesics helps to understand the spoken message

(continued)

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Table 2.2 (continued)

Author	Definition	Role of Visuals
Field (2008)	A process of interpreting a speech signal using linguistic and outside knowledge (p. 216)	Situational and paralinguistic cues provide context and help to interpret the acoustic input
Vandergrift & Goh (2012)	A process of constructing meaning from an acoustic signal by integrating cognitive processes, metacognitive strategies, and linguistic, pragmatic, discourse, and prior knowledge (pp. 33-34)	No role allocated to visuals
Rost (2016)	The integration of neurological, linguistic, semantic, and pragmatic processing (p. 50)	The interpretation of nonverbal cues as well as visual aids is part of the linguistic processes

ILA = International Listening Association

As can be seen in Table 2.2, the decoding of visual information is another aspect of the L2 listening skill that has evolved over time. Beginning with Richards in the 1980s, visual decoding has been viewed as part of the definition of the listening skill. More contemporary applied linguistics scholars mostly agree that the decoding of visual information is an essential part of the listening process. Some scholars consider visual decoding to be a part of linguistic processing (Rost, 2016) while others view it as a separate sub-process within listening (Flowerdew & Miller, 2005). Vandergrift and Goh (2012) did not explicitly allocate a role to visuals in their theoretical model. However, in the discussion of listening instruction, they admitted that processing visuals during listening is authentic and helpful for comprehension (pp. 176-177).

Notably, the majority of the visual-inclusive listening definitions only account for nonverbal visual cues, or kinesics, which are believed to provide help with contextualizing the listening message and occasionally clarifying its meaning (e.g., Field, 2008). Few definitions include the processing of visuals of content-related, or exophoric, types. Rost (2016) is the only proponent of including viewing graphical or textual visuals, such as visual aids during a lecture, into the listening construct.

Summary. The review of the literature concerning the contemporary understanding of the L2 listening process with respect to the role of visuals has revealed the following. First, L2 listening is a complex cognitive process integrating lower-level operations of input decoding, lexical search, and syntactic parsing, and higher-level operations of meaning construction and discourse construction. Second, visual information of different types is generally believed to advantage L2 listeners' comprehension processes. Third, the inclusion of non-verbal and situational cues is supported by contemporary L2 listening definitions while the inclusion of content-related clues into the construct may need better theoretical justification.

TLU Domain: L2 Academic Listening

Moving to the second step in defining the construct, as depicted in Figure 2.1 on p. 24, we now turn to the description of the TLU domain of academic listening. The context-oriented approach to academic listening is introduced first. Features of academic discourse are discussed next. The nature of academic listening comprehension is addressed afterwards. A brief summary finalizes the chapter.

Contextualized approach to academic listening. L2 learners use their listening ability as they accomplish language use tasks in a real life situation, called a TLU domain (Bachman & Palmer, 2010). From the TLU perspective, the definition of a listening assessment construct should reflect authentic practices found in real-life listening situations. In this regard, some authentic contexts (e.g., a telephone conversation; listening to a radio) do not rely on interpreting visual information. For such comparatively rare situations, the definition of the listening comprehension should exclude the ability to decode visual input. On the other hand, the majority of listening

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

contexts is accompanied by the presence of visuals (e.g., a lecture, talking to a friend) and, thus, may require visual-inclusive definitions (Suvorov, 2015a).

This differential tactic to defining a listening construct is in line with the constructivist epistemology, which takes root in the work of such scholars as Dewey (1929), Vygotsky (1986), and Piaget (1980). Constructivism postulates that scientific conceptions are determined by the contextualized reality they describe and constructed by human experiences, which paves the way for multiple listening definitions customized for specific contexts. Unlike constructivism, the realist approach argues for one objective reality that can be experienced by humans in the same way (Osborne, 1996), which would promote one universal, or default, listening construct definition.

The field of second language acquisition (SLA) has also seen two competing views on language learning and teaching that reflect the distinction between realism and constructivism. They are often referred to as the cognitive and social views (Hulstijn, Young, & Ortega, 2014). The cognitive approach to SLA pursues objectivity, accepts the existence of the universal standard of human behavior, and is mostly involved with hypothesis testing by means of quantitative research methods. Inspired by Vygotsky's (1962/1986) theory of learning, the social approach considers the context of human experience to be central and requires "social and cultural contextualization" of research findings (Hulstijn et al., 2014; p. 368).

Even though the two approaches are often deemed irreconcilable, many researchers are now in favor of using a mixed approach to SLA-related scientific inquiries. Although a likely scenario, a mixed approach would not necessarily follow a pragmatist stance and blend quantitative and qualitative research methods (Creswell &

Clark, 2011). Another possibility would be to mix the epistemology, axiology, and methodology of the post-positivist research philosophy with the ontology of the constructivism philosophical stance. In such a case, a researcher would maintain distance and impartiality of data collection using quantitative research methods while admitting the existence of multiple realities and conducting data collection within each specified reality separately. Using L2 listening as an example, such an approach would employ developing and testing *a priori* hypotheses quantitatively but separately for academic listening contexts (one reality) and general listening contexts (another reality).

A socially-oriented and constructivist-friendly approach to conceptualizing the listening skill has been recently favored by scholars (Bodie, Janusik, & Välikoski, 2008). Advancing context-independent listening theories is no longer viewed as sufficient or desirable. It is strongly suggested that listening research is contextualized, with primary attention paid to business, healthcare, education, and religious contexts. Developing multiple context-specific definitions is now a priority for listening research (p. 11).

In light of this, developing a construct definition of L2 academic listening requires the understanding of academic listening contexts. The following sub-sections describe this context in terms of typical features of academic discourse, including structural, stylistic, linguistic, and visual features, and characteristics of academic lecture comprehension, including taxonomies of academic listening sub-skills and evaluation of lecture comprehension.

Features of academic discourse. According to Lynch (2011), academic contexts require listening to academic lectures, talks, and conference presentations, as well as participating in office hours, seminars, and study groups. Among these, lectures are the

most typical type of discourse (Powers, 1986; Flowerdew, 1994). A lecture is “a setting where the subject matter of a course is explained, discussed or otherwise taken up in a meeting between lecturers and students” (Mason, 1994, p. 203). Although lecture discourse is unique for every individual speaker, it has some typical structural, stylistic, lexico-grammatical, and visual characteristics.

Structural features. Lectures normally have a logical argument structure that unfolds in sequenced chunks (Hansen & Jensen, 1994). This global organization pattern facilitates learners’ understanding and helps to avoid ambiguities. According to Young (1994), the macro-structure of the lecture is comprised of phases. Young distinguished between six main phases, namely discourse structure (orienting students), conclusion (summarizing points), evaluation (endorsing information), interaction (dialoging with an audience), theory (transmitting content information), and examples (illustrating concepts). Young argued that this is a more authentic representation of lecture structure than Woods’ (1978) beginning-middle-end configuration. The phases can be signaled by discourse markers that make a transition to the following segments of discourse, such as topic markers (e.g., *Lemme start with...*) and summarizers (e.g., *to wrap up, ...*) (DeCarrico & Nattinger, 1988). Young’s phasal structure is thought generalizable to most disciplines, with some phases possibly more prominent in certain disciplines than others.

Similarly, the prominence of micro-level structural features may also differ by discipline. For example, lectures in humanities and social sciences may have more digressions and remarks than hard science lectures (Cook, 1975; Murphy & Candlin, 1979). Lectures may also differ in the number of density of *idea units*. The term was coined by Chafe (1979) and refers to bursts of language that have a single intonation

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

contour and are followed by a pause. In the L2 listening process, idea units contribute to meaning construction and discourse construction of a lecture (Field, 2013). The mean number of words per idea unit in a lecture is 11 compared to 7 in a conversation (Chafe, 1979). One may argue that the nature of idea units depends on lecturing style, with more conversational lectures comprised of fewer idea units.

Stylistic features. Several classifications of lecturing styles have been proposed. Morrison suggested categorizing lectures into *formal* and *informal* (as cited in Jordan, 1997). According to Morrison, informal lectures may be harder to comprehend than formal lectures. Another frequently cited classification distinguishes between three lecturing styles including a *reading style*, a *conversational style*, and a *rhetorical style* (Dudley-Evans & Johns, 1981). Reading-style lecturers deliver their messages by mostly reading them from the script. Conversational lectures are less dependent on notes, more informal, and more interactive. Rhetorical style is performance-oriented and characterized by frequent jokes and other digressions. Each of the three styles generates unique tone, tempo, and intonation patterns. The choice of lecturing styles may depend on several factors, including lecture purpose and cultural conventions. However, there is some evidence showing that, on average, the conversational style is becoming more predominant, particularly in North America (Benson, 1994; Dudley-Evans, 1994; Flowerdew, 1994; Mason, 1983; McDonough, 1978).

Linguistic features. Lectures are characterized by several distinctive lexicogrammatical features. As rapid auditory acts happening in real time, lectures usually have language features typical of conversational registers, such as false starts, hesitations, irregular pausing, and other disfluencies (Flowerdew, 1994). On the other hand, lectures

have fewer turn-taking structures and indirect speech acts compared to conversations.

Vocabulary in lectures is scholarly rather than colloquial, requiring a highly developed vocabulary base (Kelly, 1991; Hansen & Jensen, 1994). Lecture syntax is also thought to be more complex than in other auditory registers, primarily because of its detailed scholarly nature. As reviewed in Lynch (2011), these complexities may be aggravated by a high speech rate or accentedness of the lecturer, the level of content abstractness, and the need to understand and produce simultaneously (e.g., listening and note-taking).

Visual features. To help listeners to cope with these complexities, the academic listening TLU domain is accompanied by visual aids that predominantly include PowerPoint presentations or hand-outs with textual, graphic, or numerical information illustrating and explaining the concepts or delineating the structure of a lecture (Field, 2011; Lynch, 2011). Using PowerPoint presentations for these purposes has become a regular practice, or a default, in lecture delivery worldwide. PowerPoint slides visualize information either in a text or graphical form at the propositional level (Field, 2009). The oral signal in lectures can sometimes become redundant because some lecturers would read the information off the slides with minimal details added verbally.

In addition to content-rich visual aids, presenter or lecturer's non-verbal cues are usually accessible to academic listeners, such as facial cues, gestures, and movements (Richards, 1983). Lectures naturally produce non-verbal signals to maintain communication with the audience and convey their purpose, attitude, or feelings. Another available visual resource is what scholars call situational or environmental cues (Bejar et al., 2000; Ur, 1984). They provide generic information about the setting, speaker, and audience, such as spatial relationships or prevailing atmosphere.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

With these multiple visual sources, the domain of academic listening can be considered multimodal, where the information is received and exchanged through at least two modalities – the auditory modality and the visual modality. As suggested by Morell, Garcia, & Sanchez (2008), this multimodality may have more than two dimensions. It most likely relies on the intercourse between spoken, textual, image-based, and other non-verbal information. In this respect, the L2 academic listening skill is an integrated construct that comprises the skills of listening, viewing, and reading.

Academic listening comprehension. It is generally held that academic listening operates on the same core processes as models for general listening comprehension do. Therefore, the processes outlined by Field (2013) sustain L2 academic listening as well. The academic listener still has to employ their lower-level processes (i.e., input decoding, lexical search, and syntactic parsing) and higher-level processes (i.e., meaning construction and discourse construction) to construct a full mental model, or a comprehensive meaning, of a lecture. What is added to these processes is an additional load due to regular complexities of academic discourse as well as the use of context-specific sub-skills imposed on the academic listener.

Taxonomies of academic listening sub-skills. Richards (1983) was the first to propose a complete set of L2 academic listening micro-skills. Broadly speaking, they include the abilities to (a) identify lecture's topic, purpose, scope, and discourse, (b) cope with differing lecture styles, modes, registers, accents, and speeds, and (c) infer or recognize relationships, key lexical items, and speaker's non-verbal signals. This taxonomy takes account of the distinct features of the academic discourse described in the previous section. Most of these micro-skills were further evaluated for their

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

importance in academic listening contexts by Power's (1985) survey of lecturers in US universities and found to be relevant. Several additional skills were added to Power's survey, including retaining information through and retrieving information from note-taking. They were also judged as highly important for L2 academic listening.

Note-taking is now considered a part of the L2 academic listening comprehension process (Flowerdew, 1994). It can be defined as condensing and paraphrasing the auditory and visual input in a written form (adapted from Aiken, Thomas, & Shennum, 1975). As an integral part of authentic lectures, note-taking assists listeners in understanding longer, informationally dense stretches of discourse and likely help to integrate information from different sources, such as PowerPoint slides, textbooks, and the spoken input. Some studies have shown a positive correlation between the completeness of notes and listening comprehension (e.g., Dunkel, 1988). Other studies found no effect of note-taking on listening comprehension scores. For example, in Carrell, Dunkel, and Mollaun's study (2002), listening scores were positively related to note-taking for lectures on arts and humanities topics but not physical sciences. What is more, note-taking did not cause significant differences in listening scores for longer passages (about 5 minutes) while exerting a positive effect on the comprehension of shorter stimuli (about 2 minutes). In terms of Field's (2013) L2 listening model, note-taking has potential to facilitate the process of meaning construction and more so at the level of discourse construction rather than lower-level processes.

Evaluation of lecture comprehension. To find an appropriate way to measure L2 lecture comprehension, it is wise to consult practices regarding how lectures are evaluated in authentic contexts. One of Powers' (1985) survey questions concerned the

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

importance of various listening activities by discipline. As judged by lecturers, the three most important activities that related to lecture assessment were (a) identifying major ideas and themes, (b) identifying relationships among main ideas, and (c) identifying supporting ideas and examples. Further, the participants evaluated the appropriateness of various specific tasks for measuring lecture comprehension. Among others, inference questions, detail questions, and selected-response tasks were the most appropriate.

Similar results were generated by Rosenfeld, Leung, and Oltman's study (2001). The study compared academic importance ratings for different reading, writing, speaking, and listening tasks, as judged by undergraduate and graduate faculty and students in American and Canadian universities. All the listening tasks included in the survey were deemed either important or very important by both faculty and students. The highest ratings were given to the abilities to understand factual information and details, identify the main ideas and their supporting information, understand important related terminology, and make appropriate inferences based on the information in a lecture.

The results of these studies largely reflect the current approach to measuring L2 academic listening. It is believed that L2 academic listening tests should assess the ability to understand both *explicit* and *implicit* information, often referred to as *local* and *global* meaning (Aryadoust, 2013; Hansen & Jensen, 1994). Powers' supporting ideas and examples tasks and Rosenfeld et al.'s factual information and details require identifying *local* meaning. Such tasks target interpretations mostly at levels of lexical search, syntactic parsing, and sometimes even input decoding. Powers and Rosenfeld et al.'s main ideas and inferences tasks require identifying *global* meaning. Though triggered by lower-level processes, such tasks mostly rely on the processes of meaning construction

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

and, to some extent, discourse construction (Field, 2013). Thus, it may be expected that lower proficiency test-takers will experience more difficulties with *global* questions and be more comfortable with *local* questions (Becker, 2016).

Summary. To reflect the realities of the TLU domain, the L2 academic listening construct should include specific skills that are not normally part of the general L2 listening framework, such as managing the comprehension of an informationally dense lecture-like stimuli. Lecture comprehension may depend on processing content-rich visuals and taking notes. Therefore, the definition including the processing of the spoken lecture along with the lecturer's non-verbal cues only would not be fully reflective of the demands of academic listening. In terms of representing the academic TLU domain, the presence of content-rich visual aids in the construct, including text and graphic, is warranted. In light of this, the existing general and academic definitions of L2 listening may be deficient (except for Rost, 2016). This lends support to re-working definition of L2 academic listening as follows: The active process of receiving and constructing meaning from the spoken lecture input, the lecturer's non-verbal cues, situational cues, and content-rich visual aids with the help of note-taking.

Comparative Empirical Studies

Having defined the L2 academic listening construct, we move to the stage of construct justification, as depicted in Figure 2.1 on p. 24 of this literature review. Reflecting the way to back the explanation assumption within an interpretive argument (Chapelle et al., 2008), the next logical step to justify the use of the visual-inclusive L2 academic listening construct would be to show that listening difficulty is systematically influenced by the presence of visuals in keeping with theoretical expectations. According

to the theory, visuals are generally expected to decrease listening difficulty, at least for middle- and higher-level listeners. Numerous studies tested this expectation by comparing the difficulty of video-based versus audio-only listening tests. An overview of these studies is the purpose of this section.

Video effect. Studies into the effects of video-based visuals on L2 listening difficulty have been reviewed. The criteria for selecting empirical studies for review included the following: (a) the effects of videos on L2 listening test scores in an academic or general listening context were the object of exploration, (b) the study was published in a peer-reviewed journal, and (c) inferential statistical methods were used to arrive at conclusions. Overall, 15 studies were selected, with two of them reporting on two separate experiments each (i.e., Baltova, 1994; Lesnov, 2017), and another two having distinct findings for different proficiency levels (i.e., Latifi, Tavakoli, & A'lipour, 2013; Parry & Meredith, 1984). These studies were organized into three groups based on their findings, namely positive, negative, and neutral effects of videos on L2 listening comprehension.

Positive video effect. Firstly, there have been eight studies documenting that test-takers' performance was positively influenced by the use of video-enhanced listening passages (i.e., Baltova, 1994; Latifi, Tavakoli, & A'lipour, 2013; Lee & Lee, 2015; Lesnov, 2017; Parry & Meredith, 1984; Shin, 1998; Sueyoshi & Hardison, 2005; Wagner, 2010b). These eight studies are displayed in Table 2.3 alphabetically. The first column references the studies by author's name and publication year. The following four columns describe participants, listening instruments, procedures, and findings.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Information about participants is organized into (a) number, context, and educational level, (b) L2 proficiency level, (c) age, and (d) gender. The number of participants ranged from 42 to 202 with about half using samples of around 45 people and the other half including more than 150 participants. Six studies targeted the English language while the other two French and Spanish. All but one targeted college students. Participants' English proficiency levels varied from beginner to advanced. Six studies did not report participants' gender profile, with the other two being female-dominated.

Listening instruments are described in terms of (a) listening type, namely interactive or monologic, and text type, such as a lecture or a documentary; (b) authenticity of the input, namely authentic (unscripted), semi-authentic (semi-scripted), and inauthentic (scripted); (c) type of video-based visuals, namely content or context; (d) type of comprehension questions, such as open-ended or multiple choice; (e) type of scoring, namely polytomous or dichotomous; and (f) instrument's internal consistency index. Measurement instruments mostly comprised dialogic, inauthentic materials related to general listening. Only one study used authentic materials (lectures). Half of the studies targeted academic listening using monologic lectures. Most studies reported using both context and content visual information, with the latter ranging from movie context through animated stories to lecture-related visual aids. Dichotomously-scored multiple-choice items were used predominantly by the researchers, with only two studies employing polytomously-scored open-ended questions. Five of the instruments contained about 20 items; the other three had from 30 to 60 items. All studies but one reported internal consistency indices, which were 0.73 and higher.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Table 2.3

Positive Video Effect on Comprehension: Comparative Studies (n=8)

Study	Participants	Listening instrument	Procedures	Findings
Baltova (1994) Exper. 1	a. 53 French as L2 learners in Canada, secondary school b. intermediate c. age of grade 8 d. not reported	a. general: mix of monolog and dialog (movie) b. inauthentic: Scripted c. context visuals (situational, kinesic) content visuals (movie content) d. multiple choice, $k = 16$ e. dichotomous scoring f. not reported	a. condition (sound-only, video-and-sound, silent viewing) b. item pre-view: No c. note-taking: No d. Independent t tests	<ul style="list-style-type: none"> The test was significantly easier under the video-and-sound condition than the sound-only condition No difference between the video-and-sound and silent viewing conditions
Latifi, Tavakoli, & A'lipour (2013)	a. 48 EFL students in Iran b. intermediate and advanced c. 15-28 years old d. not reported	a. general: mix of monolog and dialog (documentary) b. inauthentic: Scripted c. context visuals (situational, kinesic) d. multiple-choice, $k = 20$ e. dichotomous scoring f. r (Cronbach's) = 0.81	a. mode of presentation (audio vs video), proficiency level (intermediate, advanced) b. item pre-view: Yes c. note-taking: No d. ANOVA	<ul style="list-style-type: none"> The test was significantly easier for test-takers of intermediate proficiency
Lee & Lee (2015)	a. 177 EFL learners in Taiwan, undergraduate b. Intermediate c. 20-22 years old d. 136 female, 41 male	a. general: Monologic (fiction story) b. inauthentic: Scripted c. content visuals (animated stories) d. multiple choice, $k = 15$ e. dichotomous scoring f. r (Cronbach's) = 0.93	a. treatment (audio with simultaneous story script reading vs audiovisual without script reading) b. item pre-view: No c. note-taking: No d. Independent t tests	<ul style="list-style-type: none"> The test was significantly easier under the audiovisual condition than the audio-with-script-reading condition
Lesnov (2017) Exper. 2	a. 44 ESL students in the US, mostly Arabic and Chinese b. upper intermediate; TOEFL iBT: 57-68 out of 120 c. 18-25 years old d. not reported	a. academic: Monologic (lecture, presentation) b. authentic c. context visuals (situational, kinesic) content visuals (text, graphic, images) d. multiple-choice, $k = 20$ e. dichotomous scoring f. dependability = 0.75	a. mode of presentation (audio-video vs video-only); amount of content clues in a video (63%, 29%, 0%, 0%) b. item pre-view: Yes c. not-taking: Yes d. Independent t test	<ul style="list-style-type: none"> Testlet with the highest amount of content clues was easier in the audio-video condition
Parry & Meredith (1984)	a. 178 Spanish learners (native English speakers) in US, undergraduate b. Beginner, Intermediate, Advanced c. not reported d. nor reported	a. general: Dialogic (conversation) b. inauthentic: Simulated c. context visuals (situational, kinesic) d. multiple choice, $k = 60$ e. dichotomous scoring f. r (Cronbach's) in range of 0.79-0.94 for different proficiency groups and test versions	a. treatment (audiotape vs videotape) b. item pre-view: No c. note-taking: No d. Paired t tests for each proficiency group	<ul style="list-style-type: none"> The test was significantly easier under the video condition for beginner and intermediate learners

(continued)

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Table 2.3 (continued)

Study	Participants	Listening instrument	Procedures	Findings
Shin (1998)	<p>a. 83 ESL students in US, undergraduate, graduate, and unclassified; Asian: Chinese, Indonesian, Korean, Malaysian, Taiwanese, Japanese, Pakistani, Arabic</p> <p>b. Measured on an academic test: $M \approx 10.5$ out of 18, $SD \approx 2.5$; TOEFL (paper-based) ≥ 500</p> <p>c. not reported</p> <p>d. not reported</p>	<p>a. academic: Monologic (lecture)</p> <p>b. authentic for video lectures; inauthentic for audio-only lecture versions: Modified language input (long silences, dysfluency markers, ungrammatical points omitted)</p> <p>c. context visuals (situational, kinesic) content visuals (blackboard notes)</p> <p>d. open-ended, $k = 32$</p> <p>e. polytomous (partial) scoring</p> <p>f. r (Cronbach's) = 0.89 for video and 0.86 for audio-only versions</p>	<p>a. channel of presentation (audio-channeled vs video-channeled with pre-viewed printed background information)</p> <p>b. item pre-view: No</p> <p>c. note-taking: Yes</p> <p>d. Independent t test</p>	<ul style="list-style-type: none"> The test was significantly easier under the video condition (with pre-viewed background information) than for audio-channeled condition (without pre-viewed background information)
Sueyoshi & Hardison (2005)	<p>a. 42 ESL students in US, mostly Korean</p> <p>b. lower intermediate, advanced</p> <p>c. 18-27 years old</p> <p>d. 29 female, 13 male</p>	<p>a. academic: Monologic (lecture)</p> <p>b. semi-authentic: Semi-scripted</p> <p>c. context visuals (situational, kinesic)</p> <p>d. multiple choice, $k = 20$</p> <p>e. dichotomous scoring</p> <p>f. r (KR-20) = 0.73</p>	<p>a. stimulus condition (audio-video-gesture-face vs audio-video-face vs audio-only), proficiency level (lower, higher)</p> <p>b. item pre-view: No</p> <p>c. note-taking: No</p> <p>d. ANOVA</p>	<ul style="list-style-type: none"> The test was significantly easier under the two audio-video conditions than under the audio-only condition No difference in test difficulty under the audio-video gesture-face versus face conditions The audio-video-face stimulus was the easiest for advanced test-takers The audio-video-face-gesture stimulus was the easiest for lower intermediate test-takers
Wagner (2010b)	<p>a. 202 ESL learners in US, mostly Spanish, Japanese, Korean, Chinese, and French</p> <p>b. Beginner, Intermediate, Advanced</p> <p>c. 18-60 years old</p> <p>d. not reported</p>	<p>a. general: Dialogic (conversation); academic: Monologic (lecture)</p> <p>b. inauthentic: Semi-scripted</p> <p>c. context visuals (situational, kinesic) content visuals (photographs)</p> <p>d. multiple choice and open-ended, $k = 40$</p> <p>e. dichotomous and polytomous (partial) scoring</p> <p>f. r (Cronbach's) = 0.88</p>	<p>a. treatment (audio-only vs video), text type (dialog vs lecture)</p> <p>b. item pre-view: Yes</p> <p>c. note-taking: Not reported</p> <p>d. MANCOVA, ANCOVA</p>	<ul style="list-style-type: none"> The test was significantly easier under the video condition for both dialog and lecture tasks 12 items were significantly easier under the video condition (presumably due to the presence of non-verbal cues and content-related photographs) 1 item was significantly harder under the video condition (no reason offered)

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Next, procedures are reviewed in regards to (a) independent variables in the study and their operationalizations; (b) item preview availability; (c) note-taking availability; and (d) statistical analysis used. All the studies were analyzed for the effect of delivery mode, referred to by different studies as stimulus condition, mode or channel of presentation, or treatment. It was mostly operationalized as either audio-only or audio-video. Three studies allowed for pre-viewed background information and differing amounts of gestures in videos. In one study, the audio-only condition was tainted by the availability of a listening script. Half of the studies included additional variables, such as proficiency level, amount of content clues in a video, and text type. Notably, interactions between delivery mode and proficiency level were analyzed only in studies involved with context visuals. Five studies allowed pre-reviewing items while the other three did not. Only two experiments, which both targeted academic listening, offered a note-taking opportunity for test-takers. The statistics used in the studies were mainly independent *t*-tests, with occasional instances of analysis of variance (ANOVA).

Finally, regarding the findings, the first six studies found listening tests to be significantly easier under an audio-video condition than an audio-only condition for students of intermediate L2 proficiency. Sueyoshi and Hardison reported a positive effect of context-related videos regardless of proficiency level (low-intermediate and advanced). They found, however, that the audio-video-face stimulus was the easiest for advanced test-takers while the audio-video-face-gesture stimulus was the easiest for intermediate test-takers. The findings of Latifi et al.'s study showed a positive effect of context visuals only for test-takers of intermediate proficiency. In Wagner's study, the video group significantly outperformed the audio-only group overall as well as on a

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

number of individual items, both on dialog and lecture tasks. In his study, the groups were a mixture of proficiency levels ranging from beginner to advanced.

The eight described studies have strengths and weaknesses. Mostly high reliability indices for the listening instruments and adequate sample sizes can be counted as strong points. Among drawbacks, the way different kinds of visuals were accounted for should be noted. Many instruments contained mixtures of context- and content-related visuals with little or no specifications regarding the configuration of the two visual kinds in a video. Different studies likely used different configurations of context-versus-content visuals in the videos. Therefore, it is unclear which of the two kinds or which configurations caused positive effects on comprehension. In addition, few studies investigated interactions between delivery mode and proficiency level. These studies showed the potential of proficiency to impact the effect of visuals on comprehension, which signals the need to control for proficiency more thoroughly. Lastly, only one study attempted to analyze the effects of delivery mode on difficulty of individual items. Yet this analysis was largely qualitative and, thus, lacking statistical evidence.

Negative video effect. Secondly, two studies reported a negative effect of video-enhanced listening texts on test performance, as displayed in Table 2.4. This table follows the same format as Table 2.3. Both studies used mostly male Chinese or Arabic ESL students of about 20 years old. Students were at the intermediate level in one study and at the advanced level in the other. Instruments in both studies built upon academic listening scripted materials followed by 20-30 multiple-choice questions with internal consistency ranging from 0.59 to 0.70. Both studies employed videos with no content-related clues. Along with investigating the effect of delivery mode on listening comprehension, each

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

study employed an additional independent variable. It was memory capacity (i.e., low, mid, high) for one study and text type (i.e., dialog, lecture) for the other. None of the studies allowed note-taking.

Table 2.4

Negative Video Effect on Comprehension: Comparative Studies (n=2)

Study	Participants	Listening instrument	Procedures	Findings
Pusey & Lenz (2014)	a. 24 ESL students in US, Chinese and Arabic b. intermediate c. 18-24 years old d. mostly male	a. academic: Monologic (lecture) b. inauthentic: Scripted c. context visuals (situational, kinesic) d. multiple-choice, $k = 20$ e. dichotomous scoring f. r (Cronbach's) = 0.59	a. input format (audio vs video), working memory capacity (high, mid, and low) b. item pre-view: Not reported c. note-taking: Not reported d. Mann-Whitney U test	<ul style="list-style-type: none"> • The test was significantly harder under the video condition • Test-takers with low working memory capacity scored lower under the video condition
Suvorov (2009)	a. 34 non-native English speakers in US: 22 international undergraduates, 12 ESL students; mostly Chinese b. high-level c. 18-20 years old d. 9 female, 25 male	a. academic: Monologic (lecture) and dialogic (campus conversation) b. inauthentic: Scripted c. context visuals (situational, kinesic) d. multiple-choice, $k = 30$ (10 for photograph; 10 for video-mediated; 10 for audio-only part) e. dichotomous scoring f. r (KR-20) = 0.39 to 0.63, depending on test part	a. visual input type (photograph vs video-mediated vs audio-only), text type (dialog vs lecture) b. item pre-view: No c. note-taking: Yes d. ANOVA	<ul style="list-style-type: none"> • The difficulty of dialogs was not affected by the presence of video • The difficulty of lectures was harder under the video condition

Regarding the findings, Suvorov (2009) found that video-mediated lectures were harder for higher-level students than their audio-only counterparts whereas dialogs were not affected by the presence of visuals. Pusey & Lenz (2014) concluded that academic lectures were harder under the video condition, especially for test-takers having lower working memory capacity. Test-takers in this study were at an intermediate level of proficiency. It should be noted that these studies have reliability indices of lower than 0.7, which is widely considered the lowest satisfactory value for norm-referenced tests. In Suvorov's study, the internal consistency index for the video-based part of the listening

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

test was as low as 0.39. This may suggest that the instrument measured about 60% of construct-irrelevant attributes, which seriously undermines the findings.

Neutral video effect. The last set of nine studies found no difference between the scores of audio-only and audio-video groups, as reflected in Table 2.5. Again, following the same format as the previous two tables, Table 2.5 summarizes the information about participants, instruments, procedures, and findings. All but one study targeted ESL or EFL students at the undergraduate level and above, with their proficiency levels ranging from beginner to advanced. Gender profiles for most of the studies were not specified.

Regarding instruments, about half of the studies targeted academic English with the other half focusing on general English. Two of the academically-oriented instruments employed authentic lectures; the others used scripted or simulated materials. Video-based stimuli were mainly context-oriented, with only three studies reporting on using content visuals. The instruments were mostly multiple-choice tests having from 14 to 60 items. Essays, true/false questions, and short-answer open-ended questions were also occasionally used. The instruments' reliability indices ranged from 0.45 to 0.94.

Approximately half of the studies investigated additional independent variables, including text type and proficiency level. Two studies also attempted to control for video type. Suvorov's studies allocated videos to either context or content type. In Lesnov's study, videos were categorized by the amount of content-related information in videos, operationalized as the percentage of video time occupied by content clues. Note-taking and item preview were allowed in some studies but disallowed in others. Notably, note-taking was permitted in all the studies targeting academic listening. One no-effect study employed a Rasch analysis, which allowed for analyzing a video effect at the item level.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Table 2.5

Neutral Video Effect on Comprehension: Comparative Studies (n=9)

Study	Participants	Listening instrument	Procedures	Findings
Baltova (1994) Exper. 2	a. 43 French as L2 learners in Canada, secondary school b. intermediate c. age of grade 8 d. not reported	a. general: mix of monolog and dialog (movie) b. inauthentic: Scripted c. context visuals (situational, kinesic) content visuals (animated story) d. multiple choice, $k = 22$ e. dichotomous scoring f. not reported	a. condition (sound-only, video-and-sound) b. item pre-view: No c. note-taking: No d. Independent t tests	<ul style="list-style-type: none"> • No effect of condition
Batty (2015)	a. 164 EFL students in Japan, undergraduate b. 4 tiers of proficiency c. not reported d. not reported	a. general: Monologic, phone talk; general: Dialogic, conversation; academic: Monologic, lecture b. inauthentic: Scripted c. context visuals (situational, kinesic) d. multiple-choice, $k = 46$ e. dichotomous scoring f. r (Cronbach's) = 0.77	a. delivery format (audio vs video); text-type (monologue, conversation, academic lecture), proficiency level (4 tiers) b. item pre-view: Yes c. note-taking: Not reported d. Multi-faceted Rasch analysis	<ul style="list-style-type: none"> • No effect of delivery format • No interaction between format and text-type and proficiency level • Two items easier under in video condition • Two items harder in the video condition
Cubilo & Winke (2013)	a. 40 non-native English speakers in US: 10 international graduates and undergraduates, 30 ESL students b. intermediate to advanced c. 18-21 years old d. 23 female, 17 male	a. academic: Monologic (lecture) b. inauthentic: Scripted c. context visuals (situational, kinesic) d. essay, $k = 2$ e. polytomous scoring f. r (inter-rater) = 0.82 for essay 1 and 0.62 for essay 2	a. presentation mode (audio-only/still-picture-based vs video-based) b. item pre-view: No c. note-taking: Yes d. Paired-samples t test	<ul style="list-style-type: none"> • No effect of presentation mode for scores on essay content, organization, vocabulary, and mechanics • Significantly higher scores on essay language use under the video condition
Gruba (1993)	a. 91 ESL students in the US, mostly Chinese, Korean, Vietnamese, and Spanish graduate and undergraduates b. advanced c. not reported d. not reported	a. academic: Monologic (lecture) b. inauthentic: Simulated c. context visuals (situational, kinesic) d. multiple choice and true/false, $k = 14$ e. dichotomous scoring f. r (Cronbach's) = 0.45	a. presentation mode (audio-mediated vs video-mediated) b. item pre-view: Not reported c. note-taking: Not-reported d. Paired-samples t test	<ul style="list-style-type: none"> • No effect of presentation mode
Latifi, Tavakoli, & A'lipour (2013)	a. 48 EFL students in Iran b. intermediate and advanced c. 15-28 years old d. not reported	a. general: mix of monolog and dialog (documentary) b. inauthentic: Scripted c. context visuals (situational, kinesic) d. multiple-choice, $k = 20$ e. dichotomous scoring f. r (Cronbach's) = 0.81	a. presentation mode (audio vs video), proficiency (intermediate, advanced) b. item pre-view: Yes c. note-taking: No d. ANOVA	<ul style="list-style-type: none"> • No effect of presentation mode for test-takers of advanced proficiency

(continued)

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Table 2.5 (continued)

Study	Participants	Listening instrument	Procedures	Findings
Lesnov (2017) Exper. 1	a. 16 ESL students in the US, mostly Arabic and Chinese b. low intermediate TOEFL iBT: 32-44 out of 120 c. 18-25 years old d. not reported	a. academic: Monologic (lecture, presentation) b. authentic c. context visuals (situational, kinesic) content visuals (text, graphic, images) d. multiple-choice, $k = 20$ e. dichotomous scoring f. dependability index of 0.75	a. presentation mode (audio-video vs video-only); amount of content clues in a video (30%, 11%, 0%, 0%) b. item pre-view: Yes c. note-taking: Yes d. one-way ANOVA	<ul style="list-style-type: none"> • No effect of presentation mode, regardless of the amount of content clues in videos
Londe (2009)	a. 101 undergraduate and graduate international students in US, diverse cultural backgrounds b. mid-high to high c. 19-28 years old d. not reported	a. academic: Monologic (lecture) b. inauthentic: Simulated c. context visuals (situational, kinesic) d. open-ended, $k = 11$ e. polytomous (partial) scoring f. not reported	a. delivery format (video-based talking head, video-based full body, audio-only) b. item pre-view: No c. note-taking: Yes d. ANOVA	<ul style="list-style-type: none"> • No effect of delivery format
Parry & Meredith (1984)	a. 178 Spanish learners (native English speakers) in US, undergraduate b. Beginner, Intermediate, Advanced c. not reported d. nor reported	a. general: Dialogic (conversation) b. inauthentic: Simulated c. context visuals (situational, kinesic) d. multiple choice, $k = 60$ e. dichotomous scoring f. r (Cronbach's) in range of 0.79 – 0.94 for different proficiency groups and test versions	a. treatment (audiotape vs videotape) b. item pre-view: No c. note-taking: No d. Paired t tests for each proficiency group	<ul style="list-style-type: none"> • No effect of treatment for advanced learners
Suvorov (2013; 2015b);	a. 121 undergraduate and graduate international students in US and ESL students, mostly Chinese and Korean b. lower level, TOEFL iBT < 105 (out of 120) and higher level, TOEFL iBT > 111 c. 18-35 years old d. 60 female, 56 male, 6 unreported	a. academic: Monologic (lecture) b. authentic c. context visuals (situational, kinesic); content visuals (images, drawings) d. multiple choice, $k = 30$ e. dichotomous scoring f. r (Cronbach's) = in range of 0.63-0.72 for different test versions	a. delivery mode (video-based vs audio-based), video type (context vs content) b. item pre-view: No c. note-taking: Yes d. Independent t test; paired samples t tests	<ul style="list-style-type: none"> • No effect of delivery mode • No effect of video type

Most of the no-effect studies were conducted on test-takers of relatively high language proficiency with the use of context visuals (Batty, 2015; Cubilo & Wilke, 2013; Gruba, 1993; Latifi et al., 2013; Londe, 2009; Parry & Meredith, 1984; Suvorov, 2013; 2015b). This may indicate that context visuals do not benefit proficient L2 listeners.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Batty also found no interaction between different proficiency levels (four tiers of proficiency) and test-takers' performance. Baltova (1994) and Lesnov (2017) used videos that combined both context and content visuals in tests for intermediate and lower-intermediate test-takers respectively. However, at least in Lesnov's study, the amount of content visuals in videos was much lower than the amount of context visuals, making videos predominantly context-oriented. In this respect, the results were similar to the findings of the other studies. In addition to the conclusion of no difference in terms of participants' performance, Cubilo & Winke showed that videos distracted test-takers from the note-taking process.

Similar to the previous sets of findings, these studies have noticeable weaknesses. Three studies reported low or no reliability. Most of the studies did not account for video effect. Suvorov's attempt to classify videos using the context-vs-content dichotomy proved unsuccessful, by the author's admission. Finally, the role of video for individual item difficulties was addressed by only one study, which still yielded conflicting findings.

Reasons for mixed findings. The conflicting findings about the effect of videos can be explained by the lack of homogeneity among the reviewed empirical studies. As can be seen from Tables 2.2-2.4, the studies differed in terms of the following: test-takers' proficiency (intermediate, advanced) and cultural background (Asian, European, Arabic), video type (context, content), listening type (general, academic), item format (multiple choice, essay, short answer) and scoring (dichotomous, polytomous), quality of listening instruments (low, high, or unreported reliability), and administration procedures (item pre-view, note-taking). The variables of video type, listening type, and item format largely determined different operationalizations of L2 listening assessment constructs in

the reviewed studies. Among these, video type seems to be one of the most influential variables for research outcomes since it may largely determine test-takers' viewing behavior (Suvorov, 2015a) and helpfulness of video for comprehension (Lesnov, 2017). Video type, viewing behavior, and video helpfulness for item comprehension are discussed below as potential reasons for the mixed findings.

Video type. One plausible explanation for the conflicting findings may be the failure to control for video type. The distinction between video types was rarely the object of investigation in the reviewed studies. However, the reported procedures often imply that videos either were context-oriented or kinesic (e.g., Batty, 2015; Lee & Lee, 2015; Suvorov, 2009), or combined both context and content features to an unspecified degree (e.g., Baltova, 1994; Shin, 1998; Wagner, 2010b). Although both cohorts generated contradictory results, there are some noticeable trends. Studies that employed videos containing exclusively context features ($n = 10$), produced either negative or neutral effects, with three exceptions (Latifi et al., 2013; Parry & Meredith, 1984; Sueyoshi & Hardison, 2005). Notably, these three studies showed positive effects of videos for test-takers of beginner to intermediate proficiency while the rest of the context visuals studies dealt with higher proficiency test-takers (Gruba, 1993; Londe, 2009; Suvorov, 2009; 2013; 2015b).

In contrast, studies into videos containing content-related clues ($n = 6$) generally led to positive effects of listening comprehension, with two exceptions (Baltova, 1994; Suvorov, 2013; 2015b). These exceptions might have been caused by different amounts or helpfulness of content-related clues in the videos. For instance, Baltova (1994) and Suvorov (2013; 2015b) might have used lesser amounts of content clues than other

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

studies in the content visuals cohort or, perhaps, the degrees to which visuals were helpful for answering comprehension questions could be lower. These issues urge future studies to take account of test-takers' proficiency levels and video type, the latter begging a more meaningful taxonomy. A factorial ANOVA with at least two factors or more sophisticated methods based on item response theory may be particularly useful for investigating interactions between proficiency and video type. Only six out of the fifteen reviewed studies made use of such methods, with only two of them having video type as an independent variable.

Few studies controlled for the effects of video type on L2 listening comprehension explicitly, with using video type as a factor (Lesnov, 2017; Suvorov, 2013; 2015b). Suvorov investigated videos that were related to either content or context type. No impact of video type on test-takers' performance was found. The failure to find the effect of video type may have been due to the overlap between the context and content categories. Therefore, Suvorov suggested considering other dimensions of videos such as the degree of semantic congruity between audio and video inputs, rhetorical structure, and discourse type.

Lesnov's study used the amount of content clues as a basis for classifying videos into types. The study was conducted with 44 higher-level ESL students and investigated the effect of videos, each with different amounts of content-related clues. There were four videos, each containing content-related clues (text, graphic, photographs, or a combination) as well as context-related kinesic and situational cues. The amount of content-related visuals differed for each of the videos. The effect of visuals in only one video, which displayed textual visuals for 63% of the overall video time, was detected.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Other videos displayed the content-related cues for only about 30% of video lengths and exerted no significant effect on testlet difficulty. The researcher concluded that content-rich visuals may decrease listening comprehension difficulty for high-intermediate listeners. The study's major limitation was a low capacity of the listening instrument. The effects of each video type were analyzed based on students' performance on just one testlet, or six multiple-choice times. Future studies could adopt and augment the new way of classifying videos afforded by this study while employing more testlets and items to investigate video type.

Another trend was determined by interactions between video type and test-takers' proficiency. It seems that facilitative effects were often present for lower-level learners who viewed context-related visuals (i.e., Latifi et al., 2013; Parry & Meredith, 1984; Sueyoshi & Hardison, 2005) but largely absent for higher-level learners. Similarly, about half of the studies that employed content-related visuals for lower-level test-takers found facilitative effects on listening comprehension (i.e., Baltova, 1994; Lee & Lee; Shin, 1998) while others found no effect (i.e., Baltova, 1994; Lesnov, 2017). In addition, the majority of studies with context-related visuals failed to find a facilitative effect on higher-level learners' listening comprehension, with one exception (i.e., Sueyoshi & Hardison, 2005). Regarding content-related visuals' impact on higher-level learners, the trend is unclear.

Viewing behavior. Another possible explanation for differing findings relates to the argument that test-takers' performance depends on their viewing behavior. Viewing behavior can possibly be a function of the listener's educational background, familiarity with visual-inclusive listening, cultural expectations, and the degree of visual literacy

development among others. If, for instance, a test-taker has been taught L2 listening mostly in visual-free modes, he or she would not be used to watching while listening or efficient at interpreting visual information. This scenario could yield poorer viewing behavior.

Viewing behavior can also be a function of video type. In Suvorov's (2015a) study, test-takers spent statistically significantly more time watching content than context videos. Although the difference in watching time was not associated with test-takers' comprehension scores, this study's findings necessitate considering video type as a variable capable of impacting test-takers' viewing patterns and, thus, listening comprehension.

Item video-dependence. Another reason for mixed findings may be the failure to control for the relationship of comprehension questions to video input. All previous studies used comprehension items that could be answered from an audio input alone (i.e., video-independent items). Even though answering these items could be facilitated by content-related information (e.g., charts, graphs, tables, photos) or context-related information (e.g., kinesics) in videos, watching was not necessary. However, none of the studies specified to what extent video-based clues could lead test-takers' to the correct choices. One way to account for this would have been to quantify the degree to which performance on individual items can be helped by the video stimulus. From there, it would be reasonable to measure the proportion of items within a testlet that relied on the information from a video. This could have made comparisons of testlet and items difficulty by format more informative and moved the field forward in terms of investigating the role of individual items in video-enhanced listening test difficulty.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

This method of quantifying item video-dependence was implemented in the researcher's pilot study (Lesnov, 2017). The researcher used his own judgement to determine whether the answer to each individual item could be facilitated by viewing video-based content-related clues. If the answer to the item was provided directly or just alluded to in the video, the item was deemed video-dependent. It was found that the testlet that mostly contained video-dependent items was easier for test-takers in the video condition. On the other hand, testlets that contained few or no video-dependent items were equally difficult in audio-video and audio-only conditions. A limitation of this study was the way items were categorized. Rather than assigning a categorical yes-or-no label based on the researcher's sole perception, it would be worth having a number of ESL teachers use a semantic differential scale to give the item a number showing how close the item is to the "video-dependent" end of the scale. This quantity will help to control for the *degree* of item video-dependence as opposed to the mere presence of video-dependence.

Video effect at the item level. The effect of videos on individual items as well as on overall testlet performance can be investigated with many-facet Rasch measurement analysis (MFRM; McNamara, 1996). This statistical technique is based on Item Response Theory (IRT) and may be preferable to methods based on the Classical Test Theory (CTT) for a number of reasons. First, as an item-based technique, MFRM is less test-dependent and allows for deeper interpretations at the item level. Second, MFRM does not compare raw scores. Rather, it produces estimates of items and persons' abilities on a difficulty scale (i.e., the logit scale) that is truly continuous, which allows for more reliable conclusions. Finally, as argued by Batty (2015), MFRM would lead to "more

principled comparisons” (p. 9) between audio and video formats as it could place formats, test takers’ abilities, video types, and other variables, or facets, of interest on the same difficulty scale. This makes examining interactions between the facets more informative.

Despite the benefits of IRT, Batty (2015) seemed to be the only study that used MFRM for exploring the effect of videos on L2 listening comprehension at the item level. As part of the study, 164 EFL university students of different proficiency levels were administered a listening comprehension test in the two formats – audio and video (mainly context videos). The comprehension questions were answerable from the audio input alone. Besides comparing the difficulty of delivery formats (i.e., audio versus video) on the whole, Batty investigated the interactions between delivery format and text type (i.e., monologue, conversation, academic lecture), proficiency levels (four tiers of proficiency), and individual items. The MFRM with persons, items, and format set as primary facets, and text type and proficiency levels as dummy facets, yielded no general effect of format. Neither were interactions with text type or proficiency detected. However, the subsequent bias analysis discovered that four items displayed format-based differences in difficulty. Two items were favored under the video condition while the other two were easier in the audio format. Among possible reasons for these interactions, Batty mentioned gestures, facial expressions, and poor acting as exerting either facilitating or debilitating effect on the comprehension of the items. Even though these reasons were largely speculative, Batty’s findings show that even context-oriented videos can affect performance on originally video-independent questions. This may suggest that investigating items that do rely on video-based information may be of much promise.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Summary. There are several issues with the existing attempts to justify the inclusion of videos in L2 listening tests by comparing test-takers' performance by format. Failing to appropriately define and take into account video types, viewing behavior, and item video-dependence, recent research may have suffered from threats to internal validity. As a result, no conclusive evidence has been generated for or against using video-based visuals in L2 listening tests. To obtain credible conclusions, studies with more robust methodologies are needed that would take into account potential intervening variables of video type, viewing behavior, and item-video relationship. In addition, there have been no studies that would aim to justify or challenge the inclusion of textual visuals into the L2 academic listening construct. While some studies showed the benefits of viewing visualized text on L1 and L2 listening comprehension, these studies were not conducted with the notion of L2 academic construct in mind, and thus, can barely serve as evidence in the argument of enhancing the construct with text-based visual information.

Stakeholders' Perceptions

The perceptions and choices of stakeholders (i.e., people involved with or invested in the testing process), including test-takers, administrators, parents, teachers, and instructors, are of primary importance for building an assessment use argument (Bachman & Palmer, 2010). Gruba (2014) urged for the reworking of established language constructs by investigating learner and teacher perceptions of "new media" use in assessments. According to Gruba, the answers to the following questions would significantly assist in building new construct definitions: "What are the general perceptions [teacher perception] of the role of new media and technologies in language

assessment practices?”, “What skills do you think [learner perception] are being assessed with the use of these new media? How do you rate these tasks? What test-taking strategies do candidates employ under test conditions?” (pp. 11-12). Accordingly, stakeholders’ perceptions and opinions about the use of video-based visuals in L2 listening assessment tests could offer insight as to whether such tests appropriately reflect the intended construct. In this section, studies that investigated L2 test-takers and teachers’ perceptions about using video media in L2 listening assessments are reviewed. This has been attempted by empirically exploring how videos in listening tests influence listening authenticity, comprehension difficulty, and listeners’ motivation.

Authenticity. Li (2013) argued that two aspects of situational authenticity should be considered for justifying the inclusion of visuals into the listening assessment construct – “the degree of TLU task simulation” and “test stake-holders’ perceptions of authenticity” (p. 70).

The first aspect is often cited in defense of including videos into L2 listening tests because video-enhanced listening passages help re-create authentic contexts. At the same time, videos cannot fully simulate the inter-activeness of some TLU listening situations (e.g., participating in a study group) because videos are one-way interactions and do not allow for reciprocity (Li, 2013).

The second aspect of situational authenticity deals with stakeholders’ perceptions of authenticity. On this front, some work has been done to investigate the opinions of L2 teachers (Coniam, 2001) and test-takers (Cubilo & Winke, 2013). Coniam (2001) surveyed the opinions of 104 Hong-Kong English as a foreign language (EFL) teachers for visual effects on different aspects of listening comprehension. No difference was

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

found among the teachers' opinions about the authenticity of audio-only versus video-enhanced versions of a listening test. In contrast, in Cubilo & Winke's study (2013), six out of 28 ESL students indicated the increase in authenticity as the reason they preferred a video lecture to an audio-and-still-image lecture. These few studies are far from allowing us to draw a clear picture of stakeholders' perceptions with regards to the authenticity of visual listening comprehension, which calls for more research in the area.

Difficulty. To elicit judgments about difficulty, the majority of researchers tried to uncover test-takers' opinions about the degree to which videos contributed to the understanding of a listening message. For this, questionnaires were used for the most part, followed by interviews and verbal reports. These instruments were administered either during (verbal reports) or after the administration of a listening comprehension assessment (questionnaires, interviews).

In general, most of the studies reported that test takers' perceived videos as helpful for listening comprehension (i.e., Brett, 1997; Ockey, 2007; Progosh, 1996; Sueyoshi & Hardison, 2005; Wagner, 2008, 2010a). Questionnaire respondents in studies by Brett (1997), Progosh (1996), Sueyoshi & Hardison (2005), and Wagner (2010a) tended to agree that viewing videos made their tests easier. Specifically, test-takers pointed to facial expressions and gestures that attracted their attention and facilitated their understanding (Sueyoshi & Hardison, 2005). Interviewees' opinions in Ockey (2007) and Wagner (2008) showed a similar pattern. Four out of six ESL students in the former study found nonverbal cues helpful for comprehending a video-mediated listening passage. The latter study concluded that visual cues contributed to the processing of a listening text and the answering of comprehension items.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Suvorov (2015b) conducted the only study that investigated test takers' opinions on the helpfulness of context versus content videos. All of the 33 ESL students who took a questionnaire stated that content visual aids (i.e., power point slides) had a facilitating effect on their listening comprehension. Dissimilar to that, context videos were found distracting by 73% of respondents. Thus, it was concluded that, in L2 learners' opinion, viewing content visuals could decrease the difficulty of comprehension items.

Research into the effects of visuals on L2 listening difficulty as perceived by L2 teachers is not plentiful. It includes one study (i.e., Coniam, 2001), which is not in line with the findings on L2 learners' perceptions. Coniam (2001) administered a listening comprehension test to EFL teachers in Japan and then surveyed them about the helpfulness of video for listening comprehension. The majority of the teachers (82%) indicated that videos distracted them from a focused listening comprehension. Only 5% of participants documented positive effects of videos on their understanding, including an improved attention to the listening and a better understanding of the speaker's attitude.

Motivation. Another aspect that may be affected by the presence of visuals in listening comprehension is motivation towards listening. It was found that motivation positively correlated with listening comprehension (Vandergrift, 2005). Moreover, in Tafaghodtari & Vandergrift's (2008) study motivation was a significant predictor of L2 listening ability. Based on this, the authors claimed that motivation was a part of the "multidimensional conceptualization of the L2 listening ability construct" (p. 110). Since motivation is capable of explaining construct-relevant variance, its behavior under the visually enhanced condition can be indicative of how visuals interact with the listening construct itself, and, thus, is worth investigating.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

There have been a number of studies that explored test takers' opinions on how videos affected their motivation towards listening (i.e., Ockey, 2007; Parry & Meredith, 1984; Wagner, 2010a). In all these studies, the use of visuals was highly valued by test-takers. Specifically, participants stated that videos made the test more engaging and comfortable (Ockey, 2007), more motivating to pay attention to the video (Parry & Meredith, 1984), and more interesting (Wagner, 2010a). In Cubilo & Winke (2013), Progosh (1996) and Suvorov (2009), participants preferred visuals to their absence without mentioning reasons for this. To the author's knowledge, opinions of L2 teachers about the effects of visual cues on listeners' motivation have not been explored to date.

Summary. There is not enough evidence to make plausible conclusions as to how L2 teachers perceive the effects of visuals on L2 listening comprehension authenticity, difficulty, and motivation either in general or with regard to a particular video type. Test-takers generally perceived video-based visuals within a test as decreasing difficulty and increasing motivation. However, research into test-takers' perceptions of authenticity is scarce and calls for further investigations.

Research Gaps

The following issues have been identified as a product of the literature review. To start with, the definition of L2 academic listening comprehension remains underspecified with regards to the processing of visual information. Even though visual listening comprehension is now a common conceptualization of the L2 listening ability, it still largely fails to account for the role of content-related visuals, which are ubiquitous in academic contexts. This can explain treating content-related visuals aids as a construct-irrelevant factor by most test developers. On the other hand, the reasons behind not using

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

context-related visual information, which is generally treated as part of the L2 listening skill, remain unknown. Despite having more accessible technology and the visual-inclusive skill definitions, high-stakes L2 listening test developers seem reluctant to use any kinds of video-based visuals in their assessments.

It might be the case that test developers simply await more convincing evidence that would justify the use of visual-inclusive L2 academic listening constructs. While there is some evidence that supports the use of context-related visual information in the construct definition, it has yet to be empirically confirmed that content-related visuals, including graphical as well as textual visuals, should be part of the L2 academic listening construct. This could be accomplished by showing that the presence of both types of visuals in tests systematically affect lecture comprehension, reflecting the theoretical hypotheses. If convincing enough, this evidence is expected to trigger test developers' decisions on incorporating TLU-relevant visual types in the L2 academic listening assessment construct.

One piece of such evidence could be obtained by showing that videos have an impact on test-takers' listening comprehension and test performance. On this front, the existing research generated conflicting results, partly because it failed to account for powerful intervening variables. First, video types used in the studies either were not controlled for or had overlapping definitions. New research ideas are needed to circumvent the issues of overlapping visual elements within a video. Some ways to approach this could be to quantify the amount of video-based visual cues by visual type (e.g., amounts of content-related textual vs content-related graphic vs kinesic visuals), the amount of semantic match between the video and the audio stimulus, or the degree of

“helpfulness” of the video for answering listening comprehension questions. Such quantities could allow for a more meaningful investigation of the video effect. Second, none of the previous studies have looked at the degree to which individual comprehension questions could be answered from the video input. This information would help to analyze reasons behind item difficulty levels and determine what visual types are most helpful for comprehending individual items, possibly showing the benefit of viewing visuals of each type at the item level.

Evidence that test stakeholders endorse the presence of videos in an L2 academic listening test would further support the argument for making video-based tests. More work is required that would show that professional L2 teachers and assessment specialists support the argument. In this realm, the under-investigated areas are L2 teachers’ opinions about video effects on listening test authenticity and difficulty, and motivation towards listening in general as well as with regards to different video types. Similar insights from L2 test-takers are also needed. Though research into test-takers’ perceptions about video-based listening is more abundant, it is not yet clear if these perceptions will hold true for differing video types and viewing behaviors.

Research Questions and Hypotheses

To bridge these gaps, answers to the following two research questions were sought. These questions are stated generally, at the theoretical level. More specific sub-questions are discussed in the Methods section at the operational level.

1. *Do content-rich videos affect L2 academic listening comprehension difficulty?*

The study moved away from the context-vs-content classification of video type. Rather, it tried to pave the way for a new taxonomy of video types that would be based

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

on the amount of content clues in a video and identify videos on the continuum from *content-deficient* to *content-rich*. It was hypothesized that listening comprehension difficulty would be lower in the presence of content-rich videos and higher in the audio-only mode for higher-level students (Rost, 2016). For lower-level students this effect might be the opposite, provided that their ability was too low for multichannel processing (e.g., Mayer, 2005). It was also expected that test-takers' scores on testlets with content-rich videos would be positively related to their viewing behavior. The more time test-takers attend to video-based content clues, the more information they are expected to comprehend. This relationship might be reverse for lower-level test-takers. Lower listening ability might preclude test-takers from gleaning valuable information from visuals. Thus, more time watching videos might end up being a greater distractor for lower-level students (Gruba, 2004; Wagner, 2010a). At the item level, items that were video-dependent were hypothesized to be easier in the video condition and harder in the audio-only condition for test-takers of higher listening ability. Similarly, this hypothesis might not hold true for lower-level test-takers because of their limited language ability.

2. Do stakeholders' perceptions lend support for using content-rich videos in the L2 academic listening assessment construct?

Regarding the population of L2 learners, it was hypothesized that content-rich videos would elicit favorable opinions about the helpfulness of content-rich visuals for academic listening (Gruba, 2004; Suvorov, 2015b). This would reflect the attitudes of listeners in authentic situations, where listeners are supported and motivated by the variety of content-rich visuals (Lynch, 2011). Regarding the L2 teachers' population, a

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

pilot study showed that ESL teachers regard visuals as part of the listening skill (Lesnov, 2016). Visuals were also perceived to increase listening authenticity, listeners' comprehension and motivation regardless of their professional expertise. A similar answer was expected for content-rich visuals in the present study.

This chapter has laid the foundation work related to the understanding of the role of content-rich visual information in the assessment of L2 academic listening. The next chapter describes the methods of the study.

Chapter 3

Method

This chapter describes the methodology of the dissertation study. It is subdivided into the following sections: purpose of the study, participants, measures, procedures, research design, variables in the study, and data analyses.

Purpose of the Study

The overarching purpose of this study was to empirically support the argument for the inclusion of content-rich video-based visuals in second language (L2) academic listening tests. The study followed Kane's argument-based validity framework (Kane, 2004; 2006; 2013; Chapelle et al., 2008) including six inferences in the interpretive argument: test domain, evaluation, generalization, explanation, extrapolation, and utilization. Primarily, it was concerned with backing the explanation inference, which justifies the defined assessment construct. This was done by comparing ESL/EFL test-takers' performance on an academic listening test in the audio-only versus video-based mode, the latter exploiting content-rich videos. The audio-only mode represented a deficient visual-free listening construct while the video-based mode represented a sufficient theory-informed construct. If listening comprehension was affected by delivery mode as suggested by the theory, this was taken as evidence supporting the argument for including content-rich visuals in the construct. Additionally, ESL/EFL test-takers' and teachers' perceptions about the use of content-rich videos in L2 academic listening tests were elicited. If test-takers and teachers were in favor of visually content-rich academic listening, this further advanced the argument.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

In the process of backing the explanation inference, additional evidence was obtained for the following validity inferences: test domain (expert-based test development informed by literature review), evaluation (test scoring and test conditions control, and item analysis), and generalization (reliability and item analysis). These pieces of evidence were also included in the interpretive argument for the inclusion of content-rich visuals in the L2 academic listening assessment construct.

Participants

Participants in the study were drawn from two populations – the population of ESL/EFL learners and the population of ESL/EFL teachers. Sampling procedures and expected sample characteristics for each group are described below.

ESL/EFL learners. The population of ESL/EFL learners was defined as formal (school-affiliated) or independent learners of ESL or EFL worldwide. To form the sample of learners, ESL/EFL learners from multiple locations were contacted. Table 3.1 summarizes the information about the schools that extended their permissions to recruit English learners. The table displays schools' locations and names, their estimated student populations, and numbers of recruited participants after data screening adjustments.

Six schools and several online Facebook study groups were used for recruitment, as shown in Table 3.1. The selection of schools and platforms was motivated by (a) practical concerns, such as proximity to the researcher and personal connections, and (b) logistical concerns, such as the willingness or ability of school administrators to assist with the project. In terms of population characteristics, the Facebook groups stood out. Dissimilar to other institutions, members of Facebook groups may not have been ESL or EFL students per se. However, their affiliation with the TOEFL iBT, IELTS, or English

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

language study groups indicated the legitimacy of their English learner statuses with particular orientation to academic English learning and interest in academic English assessment. Whether independent learners or school-goers, the members of Facebook groups were assumed to fit into the defined population of ESL/EFL learners.

Table 3.1

Participants' Affiliations with Language Schools or Online Platforms

Location	Institution/Platform	Estimated student population	# recruited participants
USA	Program in Intensive English, Northern Arizona University	50	16
	English Language Center, Rochester Institute of Technology	50	2
Mexico	BA in English Language Teaching, Universidad de Sonora	100	74
Russia	BA in Linguistics, Zaoksky Christian College of Arts and Sciences	40	21
	Online English School "White Rabbit," Russia	2,000	9
	Online English School "English Dom," Russia	10,000	5
	Russian participants who did not specify the school	-	3
Facebook	Facebook groups for preparation for TOEFL and IELTS	> 300,000	13
	Total		143

The sample size of learner participants was 143. This number ensured a definitive statistical Rasch analysis (99% confidence; stability of measure within one logit; Linacre, 1994). It also helped to increase power of classical statistical analyses. To avoid exceeding over-representation of certain locations (i.e., USA-, Russia-, Mexico-, or Facebook-based), the recruitment for a subgroup was initially planned to be halted as soon as the subgroup reached 50 participants. However, due to low recruitment numbers from the other locations, about 20 more participants from Mexico were allowed.

Participants' demographics was heterogeneous. Table 3.2 describes participants' demographics with regards to location, age, and gender. USA-based participants were mostly males of about 20 years old. Fourteen of the USA-based participants were males and four were females. Mexico-based participants were similar in age, but were mostly

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

females (16 males and 58 females). While also female-dominated, Russia- and Facebook-based learners were somewhat older than participants from USA and Mexico. Overall, the sample of ESL/EFL learners was dominated by females.

Table 3.2

Learners' Location, Age, and Gender

Location	<i>n</i>	Age		Gender	
		<i>M</i>	<i>SD</i>	Males	Females
USA	18	22.00	5.95	14	4
Mexico	74	21.18	4.59	16	58
Russia	38	26.53	7.49	5	33
Facebook	13	26.71	60.3	5	8
Total	143	23.18	6.26	40	103

Table 3.3 describes learners' native tongues by location. We see that USA-based participants were mostly Chinese, which reflected the characteristics of the ESL students' population in the USA (Open Doors, 2016). There was one Korean-speaking participant and one Spanish-speaking participant. All participants from Mexico were native speakers of Spanish. Russia-based participants were mostly native speakers of Russian and some other Slavic languages. Participants from Facebook included native speakers of Arabic, Thai, French, Hindi, Spanish, and Turkish.

Table 3.3

Learners' Native Languages by Location

USA	Mexico	Russia	Facebook
Chinese (<i>n</i> = 16)	Spanish (<i>n</i> = 74)	Russian (<i>n</i> = 29)	Arabic (<i>n</i> = 5)
Korean (<i>n</i> = 1)		Ukrainian (<i>n</i> = 4)	Thai (<i>n</i> = 4)
Spanish (<i>n</i> = 1)		Unreported (<i>n</i> = 3)	French (<i>n</i> = 1)
		Romanian (<i>n</i> = 1)	Hindi (<i>n</i> = 1)
		Kazakh (<i>n</i> = 1)	Spanish (<i>n</i> = 1)
			Turkish (<i>n</i> = 1)

ESL/EFL teachers. To draw the sample of ESL/EFL teachers, members of Teaching English as a Second Language (TESOL) International Association were

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

contacted. Overall, there were 119 TESOL affiliates, or daughter associations, operating in five geographic regions, namely Asia and Oceania, Europe and Eurasia, Caribbean, Central, and South America, Africa and the Middle East, and North America (see Appendix A1). The overall number of TESOL members is about 50,000. To represent each of the five regions, a disproportionate-allocation stratified random sampling technique was used. Geographical regions served as strata, with the within-stratum sampling fraction set at 40%. For instance, out of the 16 TESOL affiliates operating in Asia and Oceania, seven affiliates (about 40%) were randomly selected. Their leadership was contacted with a request to send the members an invitation email with the link to the questionnaire. In case negative or no response was received from a selected affiliate organization, another affiliate from the nine remaining associations was randomly selected and contacted. The initially selected affiliates for each geographic region along with their contact information are presented in Appendix A2. The sample size of each of the five region-related strata were not proportionate to the population size of the same stratum, qualifying the sampling technique as disproportionate-allocation stratified sampling with equal fix-sized strata (Daniel, 2011).

The number of teacher participants was 310. Table 3.4 displays demographic information about the teachers, including professional location, or region, age, and gender. Teachers were 47 years old on average, as indicated by the total mean for age. Teachers from Central and South America were the youngest, with teachers from Europe and North America in the middle, followed by Asia and Africa. Responses were female-dominated in general.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Table 3.4

Teachers' Demographics

Region	n	Age		Gender		
		M	SD	Males	Females	Unreported
Asia and Oceania	115	49.68	12.07	32	81	2
Europe and Eurasia	36	47.03	11.37	7	29	0
Caribbean, Central, and South America	51	42.59	13.20	20	31	0
Africa and the Middle East	8	52.50	12.75	4	4	0
North America	100	46.04	12.61	21	79	0
Total	310	47.06	12.58	84	224	2

Table 3.5 shows distribution of teachers' native languages by region. Teachers from Europe, South America, and Africa were mostly native speakers of local languages. Teachers from Asia and North America were mostly native speakers of English.

Table 3.5

Teachers' Native Languages by Region

Asia and Oceania	Europe and Eurasia	Caribbean, Central, and South America	Africa and the Middle East	North America
English (n = 72)	English (n = 11)	Spanish (n = 38)	English (n = 3)	English (n = 82)
Bengali (n = 14)	Macedonian (n = 6)	English (n = 8)	Arabic (n = 2)	Chinese (n = 2)
Spanish (n = 4)	Spanish (n = 5)	Hungarian (n = 1)	Somali (n = 1)	Greek (n = 2)
Urdu (n = 4)	Czech (n = 4)	Indonesian (n = 1)	Turkish (n = 1)	Russian (n = 2)
Nepali (n = 2)	Georgian (n = 3)	Japanese (n = 1)	Ukrainian (n = 1)	Spanish (n = 2)
Telugu (n = 2)	Serbian (n = 3)	Nepali (n = 1)		Arabic (n = 1)
Unreported (n = 1)	Romanian (n = 2)	Portuguese (n = 1)		Berber (n = 1)
Chinese (n = 1)	Punjabi (n = 1)			Dutch (n = 1)
Filipino (n = 1)	Unreported (n = 1)			Georgian (n = 1)
French (n = 1)				German (n = 1)
German (n = 1)				Persian (n = 1)
Gujarati (n = 1)				Polish (n = 1)
Japanese (n = 1)				Punjabi (n = 1)
Italian (n = 1)				Portuguese (n = 1)
Kannada (n = 1)				Unreported (n = 1)
Kutchi (n = 1)				
Odia (n = 1)				
Persian (n = 1)				
Polish (n = 1)				
Tamil (n = 1)				
Vietnamese (n = 1)				
Uzbek (n = 1)				

Measures

There were four measures in the study: an academic listening comprehension (ALC) test, an academic listening proficiency test (henceforth, the anchor test), test-takers' questionnaire, and teachers' questionnaire. Each of these instruments is described in detail below.

Academic listening comprehension test. The ALC listening test contained four passages. To turn the four passages into testlets, each passage was followed by six 4-option multiple-choice questions assessing students' ability to infer main ideas ($k = 1$), to identify supporting details ($k = 3$), and to make inferences based on the listening ($k = 2$). Inference questions targeted test-takers' ability to deduce or predict relationships among concepts from the text. Because identifying main ideas can be considered a global inference, item types were balanced within testlets: Each testlet contained three inference items and three detail items. Each question was dichotomously scored (i.e., 0 or 1), setting the overall possible score to 24 points.

The development of each of the four testlets consisted of five main steps. First, four authentic video passages were searched for and found on the Internet. Second, four new videos were recorded to reflect the content and visual patterns of the original videos. Third, six comprehension items for each passage were developed, forming four testlets. Fourth, the items were trialed as part of the prototyping process (Fulcher, 2010). Finally, the test was piloted and revised (Fulcher, 2010). Each of the steps is detailed below. The last subsection describes how the relationship between individual items and visual cues was determined.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Search for authentic video lectures. Videos were searched for among academic video lectures posted on YouTube. Video passages were selected based on the number of pre-determined criteria regarding listening content. Specifically, each passage had to be authentic, representative of two scientific fields, conceptually rich, monologic, fair, and featuring a standard American accent. These criteria are summarized in Table 3.6.

Table 3.6

Listening Content Criteria for Selecting YouTube Video Passages

Criterion	Explanation
Authentic	The passage is situationally authentic. The passage is part of a genuine academic lecture or talk delivered or intended for college-level students in the USA. The lecture or talk may be of traditional format (i.e., delivered in a classroom) or of distance learning format (i.e., delivered online). The passage is a 5-to-7-minute uninterrupted video clip from the lecture or talk.
Representative of scientific fields	Two of the lectures are on the topic related to hard, or physical sciences. The other two represent soft, or social sciences.
Conceptually rich	The selected part of the lecture or talk explains an academic concept or compares two academic concepts in a given subject area. The explanation or comparison of concepts is supported by at least two major details and at least two minor details.
Monologic	The lecture or talk is mostly monologic, featuring a single speaker.
Fair	The listening content is in accord with the following principles of fairness: (a) contains few or no construct-irrelevant cognitive barriers, such as having inaccessible language difficulty, requiring specialized knowledge, or extensive background knowledge; (b) contains few or no construct-irrelevant affective barriers, such as topics or language causing strong emotions (e.g., violence, sexual behavior) or feelings (“ETS Guidelines for Fair Tests and Communications”, 2015).
Accented	Delivered by a speaker having the standard American or near-American accent, as judged by two ESL native speaking teachers.

In terms of video content, passages had to be unfamiliar to test-takers, which is why only videos with fewer than 100,000 YouTube views were targeted. In addition, videos had to contain considerable amounts of content-related graphical clues (e.g., graphs, pictures, illustrations), in addition to non-verbal cues from the speaker. If content-related clues were displayed for less than about 60% of the video time, the video was

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

excluded from the selection process (except for one video, as discussed later). This cut-off value was set because it (a) signified the predominance of content visuals in a video without overwhelming the video with visual aids, and (b) largely reflected visual configurations encountered in the respective authentic lectures on YouTube, as judged by the researcher. In the absence of more rigorous criteria for defining visual richness of a video in the assessment literature, criteria (a) and (b) were deemed reasonable for setting a 60% cut-off value. The content-related cues from the authentic lectures were planned to be reproduced in the final recorded videos in terms of both quantity and content.

According to the aforesaid criteria, four videos were selected. The videos were lectures about homeostasis, food tax, compassion, and exoplanets. Two of them were of a traditional in-class lecture type, and the other two were lectures delivered remotely for online classes. Appendix B includes information about kinds of content-related clues, lecture types, web addresses, university affiliations, and lengths of the selected video excerpts. It should be noted that the homeostasis video displayed no content-related clues. Although the lecturer referred to PowerPoint slides, they were unavailable in the video. Because the homeostasis lecture content neatly met the criteria listed in Table 3.6, it was decided that the video should be kept, and the visuals should be created from scratch, aiming to reflect the patterns learned from the other three selected video-based lectures.

Video recording of lectures. The next step of developing the ALC test was to record four new videos that would mirror the characteristics of the respective authentic videos but would have more homogeneity in terms of speech rate, time, and configurations of content-related visuals. To attain partial authenticity, the new videos were made to reflect the original videos with regards to (a) the content and language of

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

lectures, and (b) the use of content-related visuals. Borrowing information and ideas from the original videos was in keeping with Paragraph 107 “Fair Use” of the US Copyright Law (“Copyright Law of the US,” 2011).

Regarding lecture content and language, the original scripts were used with some modifications. Occasional alterations to script content or language were made to avoid ambiguity or digression, better introduce unfamiliar concepts, or add redundancy. As a rough estimate, about 80% of the original scripts was retained.

As soon as the scripts were ready, actors were recruited for video recording. ESL and freshmen composition teachers at Northern Arizona University (USA) were invited to volunteer as actors and deliver their lectures on camera. Two of the actors were males while the other two females. They all were native speakers of American English. The actors were given instructions on how to familiarize themselves with the script, a detailed outline, and the original video of the lecture (see Appendix C1). During recording, they were advised to gesture in a natural way and look at the camera’s eye while delivering a lecture. The actors stood at a podium, without walking. Occasionally reading off the outline was permissible. The actors were encouraged to use the outline sparingly while consistently maintaining eye contact with the camera. Each script was divided into three parts so that each lecture could be recorded in three separate shots. The recordings took place on different occasions, depending on the availability of the actors, but in the same room in one of the university facilities. A Canon Vixia HD Camcorder, a lightweight tripod, and a lighting kit were used for recording.

The four initial recordings were compared for speech rate and delivery styles. One recording (i.e., Exoplanets) had a noticeably slower speech rate, demanding its exclusion

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

from the study. Another American-speaking actor of the same gender was recruited, and the lecture was re-recorded with particular attention to speech rate. Later, the pilot study revealed that the lecture on food tax had sound issues. Similarly, the lecture was re-recorded with a different actor having parallel L1 and gender characteristics.

The final four videos were edited in the Mac iMovie software (iMovie, 2017). White noise was removed, and content-related visuals were added to the videos. Content-related visuals for Food Tax, Compassion, and Exoplanets were created to be imitations of content visuals from the respective original videos in terms of appearance, structure, and demonstration time. They were not exact reproductions because of adjustments made for homogeneity purposes. However, they were fairly similar to the originals. As mentioned above, the original Homeostasis video lacked content visuals. Therefore, for the contrived Homeostasis video, visuals were crafted following the pattern found in the other three videos. This pattern, along with other characteristics, is described below.

The four recorded video-based listening passages were equivalent in terms of length, speech rate, lexical complexity, and composition of content-related visuals, as evidenced by Table 3.7. They all were about 4 minutes in length, with word counts ranging from 734 to 863. As for auditory complexity, speech rates were roughly between 180 and 200 words per minute, or about four syllables per second, which is at or close to the moderately fast speech tier (Rivers, 1981). As for lexical complexity, about 90% of the words in the passages belonged to the Oxford 3000 list, defined by the Oxford Learner's Dictionary team as a list of 3000 most useful and important keywords ("The Oxford Text Checker," 2017). The proportion of academic words was calculated using the Oxford Academic Word List Checker ("The Academic Word List," 2017).

Table 3.7

Characteristics of the Listening Passages

Title	Major details of a lecture	Length		Speech rate		Lexical Complexity		Content-related visuals		
		Word count	Input length	Words per minute	Syllables per second	Oxford 3000	Academic Word List	Pictures % (#)	Graphs % (#)	Total % (#)
Homeostasis	A lecture explaining the concept of homeostasis in human bodies. (1) The control of weight by our bodies. (2) The mechanism of homeostasis. (3) The importance of controlling body temperature.	734	03:58	185	4.29	91%	6%	20.6% (3)	40.0% (5)	60.6% (8)
Food Tax	A lecture about the effect of taxes on human behavior. (1) A tobacco tax precedent: Tax rates across the US. (2) Positive effects of tobacco taxes on health (3) Can food taxes affect eating behavior?	747	04:08	180	4.02	92%	3%	20.9% (4)	39.7% (4)	60.6% (8)
Compassion	A lecture about mechanisms that make people feel compassion towards others. (1) Compassion as a function of similarity. (2) Experiment providing evidence for (1).	779	03:57	197	4.23	91%	5%	17.1% (5)	42.5% (5)	59.6% (10)
Exoplanets	A lecture about detecting the motion of exoplanets. (1) The definition of a barycenter. (2) The light Doppler effect. (3) The radial velocity method.	863	04:16	202	4.22	91%	6%	18.6% (3)	40.7% (8)	59.3% (11)

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Each passage contained 5-6% of academic vocabulary except for Food Tax, which had 3%. The difference in proportions of academic vocabulary among the passages was not considered significant. Because the passages did not differ critically on most of the discussed parameters, they were assumed to be equally difficult for listening comprehension.

Regarding content-related visual configurations, a distinction was made between pictures and graphs. Pictures were bitmap photographs or vector images that illustrated the concepts in a lecture. Diagrams, schemes, charts, or other graphical visual aids were labeled as graphs. All the videos contained approximately equal amounts of pictures (about 20%) and graphs (about 40%). Collectively, content-related visuals were displayed for about 60% of the video length of each video. Therefore, the amounts and configurations of content-related clues were assumed to be equivalent across the videos.

The construction of pictures and graphs for each video was guided by the following three principles. First, though the number of pictures in some ALC videos (i.e., Food Tax and Compassion) was equal to the number of graphs, pictures stayed on the screen for lower amount of time in each video. Pictures occupied approximately 20% of each video's length (17.1-20.9%) while graphs occupied approximately 40% of each video's length (39.7-42.5%). Overall, pictures and graphs were displayed for about 60% of the time in each video. In addition to pictures and graphs, each video displayed the lecturer's upper body (above the waist) such that the lecturer's facial expressions, hand gestures, and body movements were visually accessible. The lecturers were displayed for the whole video time in the left-hand half of the video frame while pictures and graphs

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

appeared in the right-hand of the video frame, as depicted in Figure 3.1 below. Pictures and graphs appeared one by one in a pre-determined succession for each video.

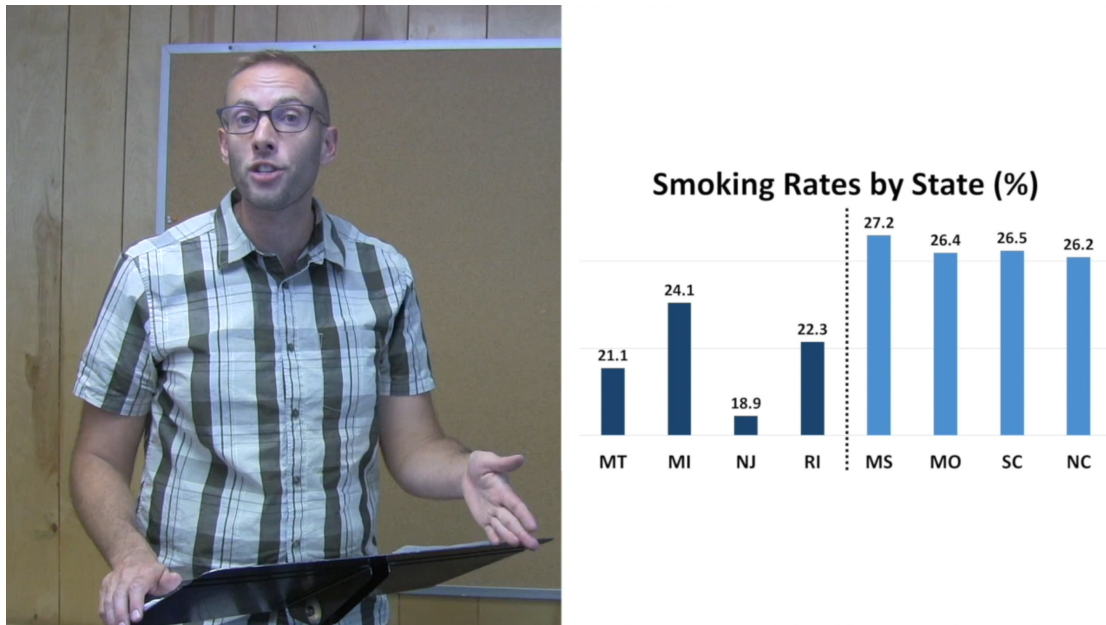


Figure 3.1. A screenshot from the content-rich video for the Taxes lecture.

This half-frame configuration was intended as a more authentic alternative to an arrangement with visual aids appearing on the whole screen, thereby systematically obstructing non-verbal cues from the lecturer. In authentic contexts, academic listeners can normally view the speaker and visuals simultaneously without obstruction. Similar to the video frame configuration in this study, lecturers in authentic contexts are oftentimes situated next to the screen with PowerPoint slides.

The second video-production principle ensured that pictures and graphs were semantically relevant to the aural input. The degree of semantic overlap between the audio and the video channels was high. Pictures and graphs in the videos contained semantically redundant information, falling into the direct category in Walma van der Molen's (2001) taxonomy of visuals. No graphs or pictures had unrelated or divergent contents. Using Bejar et al.'s (2000) terminology, all pictures were illustrating visuals.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Like pictures, some of the graphs had solely the illustrating function. Others also supplemented it with extra information or fulfilled the organizing function (see Appendix C2). For example, the lecturer in the Homeostasis video explained that the temperature at 37 degrees Celsius helped molecules in our bodies work most effectively. The corresponding graph (#8; see Appendix C2) displays an effectiveness curve with the peak at the 37-degree point, but also contains estimates for neighboring points (i.e., 35 and 39 degrees). Because these extra estimates were not explicitly stated in the audio stimulus, the graph was considered not only illustrative but also supplementing. The supplementary information in graphs provided no extra clue for answering the ALC test items. The functions of all the pictures and graphs in the study are summarized in Appendix C2.

Third, graphs and pictures were created to be easily interpretable and capable of illustrating respective verbal messages concisely and accurately, “yet with as little ink as possible” (Doumont, 2005). As a result, most of the graphs did not require a legend or written descriptions, apart from the titles, axes labels, and designations of important smaller elements. Titles, axes, and smaller elements were sparingly labeled with text and numbers. Graphs used a consistent layout across the videos, with the same font type and size, and similar positions for graphical elements.

The amount of text in graphs across the four ALC lecture videos was compared descriptively. The descriptive statistics for word counts in pictures and graphs is provided in Table 3.8. The table shows that the amount of text in video-based pictures was negligible. In contrast, word counts in graphs ranged from 2.38 to 7.75. Graphs in the Taxes video contained higher number of words than graphs in the other three videos. Exact text and word counts for both pictures and graphs are provided in Appendix C2.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Table 3.8

Word Counts in Each Video's Pictures and Graphs

	Pictures			Graphs		
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>
Homeostasis	3	0.33	0.58	5	4.60	4.22
Food Tax	4	0.25	0.50	4	7.75	4.72
Compassion	5	0.00	0.00	5	4.60	1.52
Exoplanets	3	0.33	0.58	8	2.38	1.69

Note: *n* = # of pictures or graphs per video; *M* = mean number of words per video; *SD* = standard deviation

Taking the aforesaid criteria together, a content-rich video was defined as a digital recording of a lecture that has the following properties: (a) it sequentially displays several pictures and graphs, with each positioned side-by-side with the display of a lecturer, (b) the overall display time of the pictures, graphs, and the lecturer is about 20%, 40%, and 100% of the time respectively, (c) the pictures and graphs are semantically congruous with the respective chunks of the auditory message, with the pictures fulfilling an illustrating function and the graphs serving as illustrators and/or organizers, occasionally providing some extra information not assessed by the test, and (d) the pictures and graphs are intuitive and equally easy for viewers' interpretation.

Item development. Item development was accomplished in three steps. First, the researcher drafted the items using the guidelines in Haladyna, Downing, and Rodriguez (2002) and Fulcher (2010), taking into account content, formatting, and style concerns as well as following stem- and choice-writing techniques. Second, the written items were examined by two experienced ESL teachers with item development expertise. The following problems were identified with certain items: unclear wording, implausible distractors, and questions that could be answered based on common sense. Based on this feedback, the items were revised by the researcher.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

The item writing process sought to develop two classes of items – (1) items that could be cued by video-based visual information and (2) items that could not. The video-based versions of the lectures were used as content sources for the ALC test item writing. An item's capacity to be cued by, or be answered with the help of, visual information was termed item video-dependence. In other words, if visuals were perceived helpful for getting a test item correct, the item was deemed video-dependent. Using this terminology, some of the items were written to be video-dependent while others video-independent.

The writing strategy for global video-dependent items was as follows. By definition, global items target an ability to make inferences either about relationships among ideas in a lecture or about the main idea of a lecture (Aryadoust, 2013; Hansen & Jensen, 1994). The eight video-dependent global items were written such that pictures or graphs in the video would give an *indirect* clue that helped to make a required inference. For instance, item 12, asking about the main idea of the Food Tax lecture, was designed with an assumption that the accumulation of content-related visuals in the lecture video would strongly allude to the right answer D (see Appendices C2 and D).

The writing strategy for local video-dependent items was slightly different. By definition, local items target an ability to comprehend explicit, often factual, information. The eight video-dependent local items were written such that pictures and graphs in the video would give a *direct* clue to test-takers by straightforwardly illustrating or pointing to the answer. For example, item 19 in the Exoplanets testlet asked about the meaning of a barycenter. In addition to relying on the auditory stimulus, this item could be cued by a visual of two planets balancing on a seesaw, with the balance point of the seesaw marked

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

as a barycenter (visual # 3 for Exoplanets in Appendix C2). It was assumed that this visual directly pointed to the correct option D (see Appendix D).

In contrast, design of video-independent items did not rely on video-based visual information for getting the items correct. There were no pictures or graphs that would explain, illustrate, or allude to the right answers for these items. For example, item 1 prompted test-takers to choose the correct statement based on the Homeostasis lecture. The correct answer B was not cued by any visual information in the lecture video. Similarly, video-based visual information was not sufficient for eliminating distractors A, C, or D. Therefore, item 1 was video-independent by design.

Test prototyping. Following the design and initial review processes, the items were subjected to beta prototyping, or collecting qualitative data about the items (Fulcher, 2010). Prior to launching the prototyping process, the test was posted online using the SurveyGizmo online platform (SurveyGizmo, 2017). Then, ten people were invited to take the video-based version of the test online and give their feedback about overall difficulty of the lectures, particular concerns about individual items, usability of the online interface, relevance of visual information in the videos, and other concerns they thought relevant. Six of the ten test-takers were Russia-based EFL learners at different proficiency levels. The other four were experienced second language teachers (English, Spanish). Nine test-takers took the test at their convenience at preferred locations and left their comments online upon completion of the test. One test-taker was invited to the researcher's office to take the test and to discuss it with the researcher in person.

The results of the prototyping process revealed several issues. Four items received negative feedback, with comments about implausibility of distractors and ambiguity of

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

item stems. Second, two people pointed to minor sound issues in the Food Tax video. For example, the following was pointed out: "...it was kind of hard to understand the speaker 'cuz he had to look down to his notes and his articulation and speech weren't clear."

Third, two people indicated that the 40-seconds-per-item answer constraint was insufficient and stressful. Based on this feedback, the following revisions were made: (a) the problematic items were revised in consultation with the test-takers who expressed item-related concerns, (b) the Food Tax lecture was re-recorded with a different actor, and (c) time constraints on answers were removed.

Other criticisms included high difficulty and boringness of the lectures. However, they were offset by a number of opposite comments, such as "... the lectures were interesting and educational; the interface is very moderate without distractions, understandable, intuitive..." In addition, some test-takers mentioned that the manner of lecture presentations was somewhat different from the real-life lectures primarily because the actors did not move around the classroom and did not use a more creative lecturing style. Though well taken, this criticism was left unaddressed because the present study targeted the reading lecturing style as opposed to the conversational or rhetorical style (Dudley-Evans & Johns, 1981).

Test piloting. Next, the listening test was subjected to piloting. According to Fulcher (2010), piloting is a trialing of test items with a group of about 30 people belonging to the target population. Accordingly, the four testlets were piloted with a group of 29 ESL and EFL test-takers. Sixteen were ESL learners at the university where the researcher is affiliated. The other thirteen were ESL or EFL learners affiliated with either a Russia-based online English teaching school or a Facebook TOEFL study group.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

The reliability analysis showed consistency of results across items within the test. Cronbach's alpha indices for the audio-only and video-based subsets of the test were 0.83 and 0.73 respectively. These indices showed that items worked well together and could be combined into one meaningful score for each mode. The person separation reliability index, a Rasch equivalent of internal consistency reliability ("Reliability and Separation of Measures," 2017), reached the value of 0.75. This showed that the test could adequately differentiate between at least two levels of test-takers' proficiency.

The combination of Rasch and classic analyses of items' psychometric properties was employed. The Rasch analysis was used to calculate items' infit values, which are normally taken as indicators of items' fitness for the measured construct. The classical analysis focused on two attributes of the items, including item difficulty and item discrimination, the latter estimated with point biserial correlation (Fulcher, 2010). If one or more parameters had low values, an item became a candidate for major revision. To take a further look at item functioning, a distractor analysis was run. If an item's distractor had zero attraction, it was assigned for revision.

The analyses revealed major problems with four items. In addition, the distractor analysis identified 10 implausible distractors. The revision process consisted of (a) having a consultation with an assessment expert regarding the problematic items, and (b) making revisions suggested by the expert. The revisions included rewording of both the stems and the alternatives to avoid tricky and ambiguous content, and conspicuously incorrect distractors (Fucher, 2010; Haladyna et al., 2002). The revised listening test, answer key, test specification, and lecture scripts can be found in Appendix D.

Empirical confirmation of items' visual-related designs. Once the ALC items were finalized, it was necessary to empirically confirm the researchers' decisions to classify the items as video-dependent versus video-independent. It was done using two instruments, the video-dependence questionnaire and the muted video-based ALC test.

The video-dependence survey sought to elicit three ESL/EFL learners' and three teachers' judgments about the degree of video helpfulness for answering individual comprehension questions. The main question in the survey was: "How helpful is the video-based visual information for answering this question?" (for learners) or "To what degree can the video-based visual cues help a test-taker to answer this question correctly?" (for teachers). It was answered on a 5-point scale (1 – not helpful; 5 – very helpful). The survey can be found in Appendix E.

As a result, each item in the ALC test was assigned a number from 1 to 5 three times by teachers and three times by learners. The three teachers' ratings were averaged for each item, generating an array of 24 teacher-informed item video-dependence indices. Similarly, the three learners' ratings were averaged for each item, generating an array of 24 learner-informed item video-dependence indices. The arrays of average values for each group of items were analyzed and compared using a cut-off score of 3. Values equal to or below 3.0 indicated that survey-takers did not consider visuals sufficiently helpful for answering an item. This was taken as a marker of video-independence. Values above 3.0 showed that survey-takers considered visuals sufficiently helpful for answering an item, which was taken as a marker of video-dependence.

Since teacher-informed and learner-informed item video-dependence indices measured subjective human perceptions, it was decided to supplement them with more

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

objective data about the capacity of content-rich videos to cue the ALC items. One way of determining whether video is helpful for answering comprehension questions is based on silent viewing. If test-takers are able to answer a question based on video only, it would afford additional evidence that the question is video-dependent.

To obtain such evidence, a muted version of the video-based ALC test was used. The muted version of the video-based ALC test used the same videos and questions as the regular video-based ALC test. Unlike the regular video-based ALC test, the muted version was sound-free and allowed item preview. In addition, its online interface did not have any time or navigation constraints; participants were free to replay the video or its parts as needed as well as to revisit any page.

Another three learners and another three teachers took the muted version of the video-based ALC test. The learners' and teachers' responses were analyzed separately for correct answers on each item (i.e., 1 or 0). Items with total scores lower than 2 out of 3 were flagged as lacking video-dependence. This cut-off score signified a prevalence of situations where an item was cued by videos over instances where it was not.

The video-dependence confirmation process was steered by the principle of preponderance of counterevidence. The four sources of information were examined: learners' video-dependence questionnaire ratings, teachers' video-dependence questionnaire ratings, learners' muted ALC test scores, and teachers' muted ALC test scores. An answer for the following question was sought: Is there enough evidence to exclude an item from the group with the video-dependent design? If at least three out of the four values for an item with the originally video-dependent design were below the

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

expected value, the preponderance-of-counterevidence condition was considered satisfied, and the item was assigned to the video-independent group.

Following this principle, four items were re-classified (see Appendix F). The final item classification decisions are given in Table 3.9. The ratio of video-dependent items to video-independent items in the ALC test was 14:10, or 7:5. This ratio applied to both local and global subcategories of items. Percentagewise, the test had about 58% of video-dependent items and 42% of video-independent items. Video-dependence was used as a categorical yes-or-no variable in the analyses for research questions in the study.

Table 3.9

Final Classifications of the ALC Test Items by Video-Dependence

Testlet	Video-dependent ($k = 14$)		Video-independent ($k = 10$)	
	Global ($k = 7$)	Local ($k = 7$)	Global ($k = 5$)	Local ($k = 5$)
Homeostasis	Item 4	Item 2	Item 3	Item 1
Food Tax	Item 7	Item 5	Item 6	Item 11
	Item 12	Item 8	Item 10	
Compassion	Item 13	Item 9	Item 17	Item 14
	Item 18	Item 15		Item 16
Exoplanets	Item 20	Item 19	Item 24	Item 21
	Item 23	Item 22		

Note: k = number of items

Anchor test. The anchor test measured a visual-free construct of academic English listening comprehension, similar to the construct measured by the audio-only version of the ALC test. The results of the anchor test was used to determine test-takers' academic listening proficiency. Using the ALC test for this purpose was undesirable due to possible contaminating effects of delivery mode on proficiency decisions.

The anchor test had two testlets. The two YouTube lecture clips were selected based on the same criteria that were used for selecting the four mainstream passages. The

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

credentials of the videos are provided in Appendix G. The amount of content-rich visuals was not a selection criterion because the testlets were administered in the audio-only mode. The videos were embedded in the testing software and appeared in the same form as on YouTube. Therefore, no video recording was required for the anchor testlets.

The two selected lectures were about cybersecurity and language (henceforth, *Cybersecurity and Language*). Their length and complexity characteristics are summarized in Table 3.10. The table shows that *Language* was slightly easier for test-takers than *Cybersecurity* because it was somewhat shorter, slower, and contained a smaller proportion of academic vocabulary. Such a difference is desirable for proficiency tests because it ensures that test content targets different proficiency levels.

Table 3.10

Features of the Anchor Listening Passages

Title	Major details of a lecture	Length		Speech rate		Lexical Complexity	
		Word count	Input length	Words per minute	Syllables per second	Oxford 3000	Academic Word List
Cyber-security	A lecture about the lack of trust online. (1) The essence of the problem. (2) Three types of cyber-attacks	693	04:15	163	3.71	90%	8%
Language	A lecture about how children learn their first language. (1) Basic facts about children's L1 acquisition. (2) Puzzles associated with children's L1 acquisition.	532	03:47	141	3.25	92%	5%

Items for the anchor test were developed following the guidelines in Haladyna et al. (2002) and Fulcher (2010). Reflecting the ALC test item structure, each anchor testlet contained three detail and three inference questions, the latter including one main idea

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

question. The items were subject to a thorough review by two ESL teachers and one assessment specialist, and two rounds of piloting with groups of 73 and 29 test-takers.

The first pilot generated a Cronbach's alpha value of 0.48. Problematic items were identified and revised, and one new item was added to each testlet. According to the second pilot, the reliability of the anchor test increased to 0.56, which still indicated the need to revisit and rework the items. Analyses of psychometric properties of the anchor test revealed problems with four items and one distractor. These items were revised using strategies similar to those employed for the mainstream ALC test. Appendix H contains the revised items, answer key, scripts, and table of specification for the anchor test.

Test-takers' questionnaire. A questionnaire was developed to elicit test-takers' perceptions about the effects of videos on listening comprehension. A portion of this questionnaire was administered after each of the four mainstream testlets. The questionnaire had two versions – the audio-only version, which came after each testlet in the audio-only version of the ALC test, and the video-based version, which came after the video-based versions of each ALC testlet. The audio-only version had questions about the effect of videos on listening difficulty, motivation, and authenticity, and whether videos should be used in academic tests. The video-based version also sought perceptions about viewing behavior and helpfulness of content-rich videos for answering comprehension questions. Both versions ended with four items eliciting learners' demographic information, including first language, school affiliation, age, and gender. The questionnaire design is reflected in Table 3.11. The table displays the seven above-mentioned content areas, or constructs. For each construct, the following information is provided: item type and scale, and the number of items for each version.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Table 3.11

Test-takers' Questionnaire Design

#	Content area (construct)	Item type and scale	Number of items	
			Version 1	Version 2
1	Viewing behavior	5-point Likert equivalent	-	1
2	Video effects on listening difficulty	6-point semantic differential	1	1
3	Video effects on motivation	6-point semantic differential	1	1
4	Video effects on authenticity	6-point semantic differential	1	1
5	Video helpfulness for answering questions	6-point Likert	-	1
6	Use of videos in academic tests	6-point Likert	3	3
7	Demographic information	multiple-choice open-ended	2 2	2 2
Total			10	12

Items were written in English in the form of statements, seeking the degree of test-takers' agreement. For this, a classical Likert scale, a modified no-neutral-point Likert scale (i.e., 1-Strongly Disagree, 2-Disagree, 3-Somewhat Disagree, 4 -Somewhat Agree, 5-Agree, 6-Strongly Agree), and a 6-point semantic differential scale (e.g., 1-very easy, 7-very difficult) were used. Different magnitudes of the scales were used for different questions. For example, to elicit self-ratings of viewing behavior, a 5-point Likert equivalent was more conducive because it allowed for the middle point marking a half-attentive viewing behavior. In contrast, for other questions, the neutral point was not desirable because it could invite ambivalence. A longer 6-point semantic differential scale was used to elicit more granular perceptions. Demographics-related items were either multiple-choice or open-ended.

To ensure the quality of items, two steps were taken. First, the item-writing strategies from Dornyei and Taguchi (2009), Fink (2009), and Fowler (2014) were used. They included circumventing compound and complex sentences, avoiding non-specific and loaded words, as well as pointed, double-barreled, and negatively worded questions. The items were created to sound natural and motivating to respondents. Second, initial piloting was implemented (Dornyei & Taguchi, 2009). It involved working with two

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

people unfamiliar with the specificity of the questionnaire content. One person was a native English speaker. These people took the questionnaire and provided feedback on how clearly worded, parsimonious, and cognitively heavy the items and directions were. As a result, many items were reworded and the directions were refined.

The questionnaire was piloted with the same 29 ESL/EFL test-takers that were recruited for the ALC test pilot. Since no items displayed critical problems or elicited negative comments at the review and pilot stages, the questionnaire was assumed to be in working order. However, question # 7 of the questionnaire was changed from “Which country are you from?” to “What is your first (native) language?” to better target linguistic affiliations of test-takers. The final questionnaire is found in Appendix I along with its table of specifications.

Teachers’ questionnaire. A questionnaire was developed to elicit ESL and EFL teachers’ opinions about contextualized, or immediate, perceptions about the role of content-rich videos in listening comprehension. After watching an excerpt from one of the content-rich ALC test videos, the teachers were asked about the following content areas: (1) effects of video-based content-rich visuals on academic listening comprehension difficulty, (2) motivation, and (3) authenticity, as well as (4) whether content-rich videos should be used in high-stakes listening tests. Accordingly, four multi-item scales were developed, each scale containing items relating to the corresponding content area. Items were on a 6-point Likert scale, from strongly disagree to strongly agree, with no neutral option. In addition, the questionnaire ended with seven multiple-choice and open-ended items eliciting background information (i.e., occupation, education level, L2 teaching experience, first language, age, gender, and email).

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

The process of item development and revision was identical to that of the test-takers' questionnaire (see the previous section). Unlike the final version, the initial version of the teachers' questionnaire mainly asked about general perceptions, used a five-point Likert scale and a 1-to-7 semantic differential scale, and had three extra content areas, namely role of visuals in the construct, effects of context visuals, and effects of content visuals, but did have questions about the use of content-rich videos in tests. The design of the initial questionnaire version is presented in Table 3.12.

Table 3.12

Initial Teachers' Questionnaire Design and Reliability

#	Multi-item scale (content area)	Item type	Pilot number of items	α	Final number of items
1	Role of visuals in the construct	5-point Likert	7	0.90	4
2	Effects on listening difficulty	5-point Likert	7	0.80	4
3	Effects on motivation	5-point Likert	5	0.76	4
4	Effects on authenticity	5-point Likert	4	0.50	4
5	Effects of context videos	7-point Likert	5	0.70	-
6	Effects of content videos	7-point Likert	4	0.80	-
	Demographic info	Multiple-choice	4	-	7
	Total	-	32	-	23

The initial version of the teachers' questionnaire was piloted ESL teachers in the intensive English program where the researcher was affiliated and students in M.A. in TESOL and Ph.D. in Applied Linguistics programs at the same university ($n = 42$). The analyses showed that most of the items performed as expected. Two items had low item-total correlations, and were candidates for thorough revision. Table 3.12 shows the internal consistency reliability indices for each multi-item scale after deleting these items.

During the revision process, the following modifications were made. First, all the questions were rephrased in order to tap into immediate teachers' perceptions (i.e., perceptions immediately following the watching of a content-rich video excerpt) rather than general perceptions. Being able to see content-rich videos from the ALC test is

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

believed to increase relevance of the teachers' perceptions for the argument about the inclusion of content-rich videos in the L2 academic listening construct. Second, areas 1, 5, and 6 were eliminated (see Table 3.12). These areas were no longer relevant either because they could not be transformed for immediate perceptions (area 1) or because they were redundant (areas 5 and 6). Third, a content area about the use of content-rich videos in tests was added. This area was found to be crucial for justifying the use of content-rich videos in listening assessment constructs. Fourth, the five-point Likert scale and the 1-to-7 semantic differential scale were replaced with a six-point Likert scale. This scale helped to avoid ambivalent (neutral) answers from the teachers. Fifth, the content areas were balanced in terms of the number of items, letting each be represented by four items. Redundant items were designated based on item-total correlations; items with the lowest correlations were eliminated from the content areas. Because the original area 4 (effects on authenticity) had a low number of items and a low reliability index, one item was revised and one item was added to this multi-item scale. Finally, the background section was largely revised to increase clarity. Table 3.13 below summarizes the design of the revised teachers' questionnaire. The last column of the table reflects the changes in the number of items for each content area. Appendix J contains the revised version of the teachers' questionnaire and its table of specifications.

Table 3.13

Revised Teachers' Questionnaire Design and Reliability

#	Multi-item scale (content area)	Item type	Pilot number of items	Final number of items
1	Effects on listening difficulty	4-point Likert	7	4
2	Effects on motivation	4-point Likert	5	4
3	Effects on authenticity	4-point Likert	4	4
4	Use of content-rich videos in tests	4-point Likert	-	4
	Demographic info	Multiple-choice	4	7
	Total	-	32	23

Procedures

This section describes the data collection procedures. First, it explains how the data were obtained from ESL/EFL learners. From there, it details how the data from ESL/EFL teachers were collected.

ESL/EFL learners. There were two methods of recruitment of learners, which are detailed in Table 3.14. In some schools, potential participants were given a brief in-person verbal presentation about the project by the researcher or a local teacher. Afterwards, the learners received an invitation email from the researcher with the link to the listening instruments. This recruitment method was used for schools 1, 3, and 4, as reflected in the Table 3.14. In the other schools, English learners were introduced to the research strictly by email, either directly sent from the researcher, as in 7, or forwarded by the school's administrator, as in schools 2, 5, and 6. The invitation email contained the link to the listening instruments. Regardless of the recruitment method, participants had an opportunity to leave their email addresses at the end of the test. These email addresses were later drawn into a raffle to win one of twenty \$40 USD prizes. The winners were contacted by email and given instructions as to how to receive the award.

The three test-takers' assessments, namely the ALC test, the anchor test, and the questionnaire, were combined in one academic listening assessment battery, which operated on an online testing platform run by Survey Gizmo. The battery started with the academic listening test, with section 1 of the test-takers' questionnaire appearing after each testlet, continued with the anchor test, and concluded with sections 2 and 3 of the test-takers' questionnaire. Besides the listening instruments, the battery also included

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

video instructions and text boxes for participants' feedback and emails. The time commitment for taking the assessment battery was approximately 40 minutes.

Table 3.14

Recruitment of Learners

#	Institution/Location	Recruitment method	Compensation
1	Program in Intensive English, Northern Arizona University, AZ, USA	verbal in-person recruitment by the researcher	a chance to win one of twenty \$40 prizes
2	English Language Center, Rochester Institute of Technology, NY, USA	invitation email with the link to the test	a chance to win one of twenty \$40 prizes
3	BA in English Language Teaching program, Universidad de Sonora, Hermosillo, Sonora, Mexico	verbal in-person recruitment by a local teacher and the invitation email	a chance to win one of twenty \$40 prizes
4	BA in Linguistics program, Zaoksky Christian Institute of Arts and Sciences, Zaoksky, Russia	verbal in-person recruitment by a local teacher and the invitation email	a chance to win one of twenty \$40 prizes
5	Online English School "White Rabbit," Russia	invitation email with the link to the test	a chance to win one of twenty \$40 prizes
6	Online English School "English Dom," Russia	invitation email with the link to the test	a chance to win one of twenty \$40 prizes
7	TOEFL/IELTS Study Groups on Facebook	invitation post in the group; if requested, an invitation email with the link is sent	a chance to win one of twenty \$40 prizes

The administration of the test-takers' assessment battery took place online, at each participants' convenience and preferred location. Upon following the invitation link, a participant had to view the initial video instructions. They briefly introduced the test and explained the benefits for the participant. In addition, the video instructions urged participants to check the power level of their laptops, to avoid stopping or pausing the test and reloading web pages, to remain seated at a desk while taking the test, and to be attentive listeners. The instructions also gave test-takers an opportunity of note-taking.

After listening to the instructions and electronically signing the informed consent, the test-taker was randomly assigned a number (1 or 2) by the system. If assigned the value of 1, the test-takers was administered the audio-only versions of the ALC test and

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

the test-takers' questionnaire. If assigned the value of 2, the test-taker was administered the video-based versions of the ALC test and the test-takers' questionnaire. This random assignment ensured probabilistic equivalence of the audio-only and video-based groups. The administration of the anchor test did not depend on this randomization and was identically audio-based for both groups.

To minimize the possibility of recruiting the same participant twice, the following information was monitored for each response: IP address, web browser, country, city, and postal code. These data were automatically collected and stored by SurveyGizmo. A combination of identical IP addresses, countries, and cities were considered a significant overlap between two responses. If significant overlaps were found between two or more responses, only the earliest response were used in the analysis. The data screening procedure is described in more detail in Chapter 4 of this dissertation.

The testing software automatically ran directions, listening passages, videos, and the questionnaire as well as controlled the allocation of listening time. The directions were recorded by the researcher and then included in the battery. The directions were also accompanied by respective text on the screen. The system did not allow for video replays and automatically sent the test-taker to comprehension questions upon the completion of a lecture. Test-takers were allowed to pre-view comprehension questions for two minutes before listening to each lecture. This likely helped test-takers focus more on lectures and reduced the role of memory in answering comprehension questions, thereby minimizing construct-irrelevant variance. While listening, test-takers were free to take notes if needed. The students designated their answers by a mouse click over the correct option in the tests or by typing in text for open-ended questions in the questionnaire. There were no

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

time constraints on answering test or questionnaire items. The answers were stored internally in one of the system's databases and were treated confidentially.

An approval to administer the ALC test, the anchor test, and the test-takers' questionnaire was received from the Institutional Review Board (IRB) at Northern Arizona University on September 28, 2017, with a few subsequent amendments. In line with the IRB requirements, test-takers had to sign an informed consent before taking the assessment battery. The informed consent explained the purpose, benefits, and procedures of the study, and asked test-takers for permission to use their data in the research (see Appendix D). The consent was signed online by clicking the corresponding button.

ESL/EFL teachers. The teachers' questionnaire was administered online via SurveyGizmo. First, the accessible population of TESOL-affiliated ESL/EFL teaching organizations was determined, and the sampling frame documented (see Appendix A1). From the sampling frame, 48 organizations were randomly selected using a stratified random sampling technique (see Appendix A2). The leaders of these organizations were contacted with a request to forward the invitation email to the members. The email contained a brief introduction to the research and the questionnaire link. About three weeks later, the leaders were asked to send a reminder. The questionnaire required about 15 minutes to complete. Respondents had to provide an electronic consent allowing the use of their responses for research purposes (see Appendix J). Answers from the participants and their personal information (i.e., IP addresses) were treated confidentially. Out of the overall teacher sample, 10 teachers were randomly selected to receive a \$40 USD award each. The drawing was held, and the selected teachers were given instructions on how to receive their awards.

Research Design

This study adopted the post-positivist philosophical stance. It holds that the true knowledge is observable, though all observations are fallible, and all theories are revisable (Creswell, 2013; Creswell & Clark, 2011). In line with the post-positivist epistemology, the study used the combinations of quasi-experimental and non-experimental designs and employed quantitative research methods.

To investigate the effect of videos on test-takers' performance and perceptions, the study used a quasi-experimental design with the use of nonprobability purposive sampling and a random assignment of the sample to the experimental and control groups. The treatment in this design was embedded in the instrument for the experimental group. The experimental group received an instrument with content-rich videos (treatment) while the control group received the identical instrument without videos (no treatment). Although the treatment in the study was applied somewhat unconventionally, this did not preclude the study from fulfilling two of the basic requirements of an experimental study, namely (1) having a treatment and a control group, and (2) random assignment of participants to the group (Hatch & Lazaraton, 1990). This design is depicted in Figure 3.2. The anchor test was used to control for test-takers' listening proficiency in English.

To investigate teachers' perceptions, a non-experimental one-group survey design was used ("Research Methods Knowledge Base," 2006). Effects of geographical region, L2 teaching experience, and educational level on teachers' perceptions were explored.

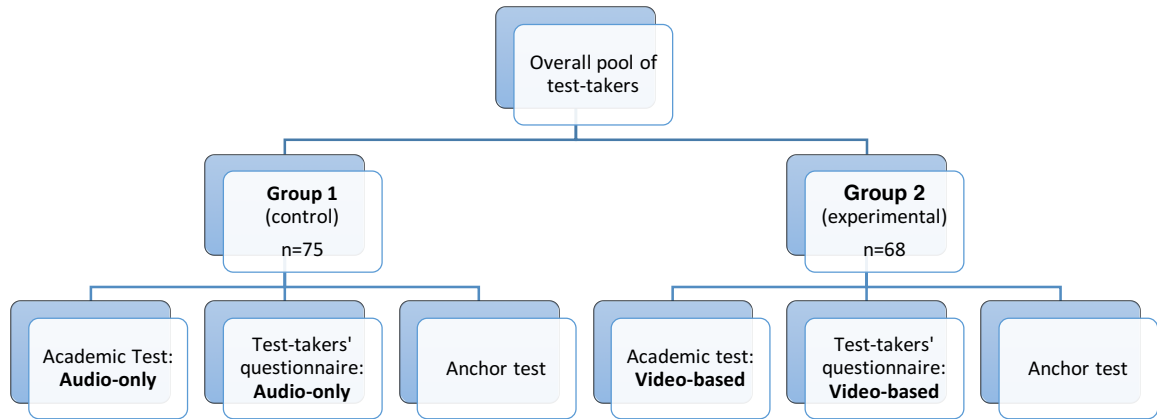


Figure 3.2. Illustration for the investigation of test-takers' behavior.

Variables in the Study

This section summarizes all the variables in the study in tables. Table 3.15 below contains the following information for each dependent variable: function, name, operationalization, measurement level, range of possible values, and research questions. The dependent variables are clustered into three groups for test performance (RQs 1.1-1.3), test-takers' perceptions (RQ 2.1), and teachers' perceptions (RQ 2.2). Overall, there were 40 dependent variables. Most of the variables were continuous, with only four being on ordinal scales (i.e., variables 4 through 7). Variables 1a and 1b had the same names but different operationalizations. Variable 1a was used in the Rasch analysis for research questions 1.1 while variable 1b was used in the classical ANOVA analysis for the same research question.

The independent variables in the study are summarized in Table 3.16. This table follows the same format as Table 3.15 above. Overall, there were 30 independent variables. All the independent variables were categorical.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Table 3.15

Dependent Variables in the Study

#	Name	Operationalization	Level	Range	RQ
1a	Difficulty at the test level	Item Rasch logit values (collectively)	Continuous	-2.06- 4.44	1.1
1b	Difficulty at the test level	Total of all ALC item scores	Continuous	1-24	1.1; 1.3
2	Difficulty of video-dependent items	Total score on ALC test video-dependent items (items 2, 4, 5, 7, 8, 9, 12, 13, 15, 18, 19, 20, 22, 23)	Continuous	1-14	1.1
3	Difficulty of video-independent items	Total score on ALC test video-independent items (items 1, 3, 6, 10, 11, 14, 16, 17, 21, 24)	Continuous	1-10	1.1
4	Difficulty of video-dependent local items	Total score on ALC test video-dependent local items (items 2, 5, 8, 9, 15, 19, 22)	Ordinal	1-7	1.1
5	Difficulty of video-dependent global items	Total score on ALC test video-dependent global items (items 4, 7, 12, 13, 18, 20, 23)	Ordinal	1-7	1.1
6	Difficulty of video-independent local items	Total score on ALC test video-independent local items (items 1, 11, 14, 16, 21)	Ordinal	1-5	1.1
7	Difficulty of video-independent global items	Total score on ALC test video-independent global items (items 3, 6, 10, 17, 24)	Ordinal	1-5	1.1
8-31	Difficulty at the item level	Item Rasch logit values (individually)	Continuous	-1.69- 0.99	1.2
32	Test-takers' viewing behavior	Sum of self-ratings of viewing behavior (item A on video version of test-takers' questionnaire; 1-5 scale) for each of the four ALC testlets	Continuous	1-20	1.3
33	Test-takers' difficulty perceptions	Sum of difficulty ratings (item 2 on test-takers' questionnaire; 1-6 scale) on each testlet	Continuous	1-24	2.1
34	Test-takers' motivation perceptions	Sum of motivation ratings (item 1 on test-takers' questionnaire; 1-6 scale) on each testlet	Continuous	1-24	2.1
35	Test-takers' authenticity perceptions	Sum of authenticity ratings (item 3 on test-takers' questionnaire; 1-6 scale) on each testlet	Continuous	1-24	2.1
36	Test-takers' opinions on using videos in tests	Sum of opinions on using content-rich videos in L2 listening tests (total score on items 4-6 on test-takers' questionnaire; 1-6 scale)	Continuous	1-18	2.1
37	Teachers' difficulty perceptions	Total score on items 1, 5, 12, 15 in teachers' questionnaire	Continuous	1-24	2.2
38	Teachers' motivation perceptions	Total score on items 2, 6, 9, 16 in teachers' questionnaire	Continuous	1-24	2.2
39	Teachers' authenticity perceptions	Total score on items 3, 7, 10, 13 in teachers' questionnaire	Continuous	1-24	2.2
40	Teachers' opinions on using videos in tests	Total score on items 4, 8, 11, 14 in teachers' questionnaire	Continuous	1-24	2.2

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Table 3.16

Independent Variables in the Study

#	Name	Operationalization	Level	Range	RQ
1	Delivery mode	Audio-only Video-based	Categorical	1-2	1.1; 1.2; 2.1
2	Test-takers' listening proficiency	Lower Higher	Categorical	1-2	1.1-2.1
3-27	Item video-dependence	Video-dependent Video-independent	Categorical	1-2	1.1; 1.2
28	Teachers' geographic region	Asia and Oceania Europe and Eurasia Caribbean, Central, and South America Africa and the Middle East North America	Categorical	1-5	2.2
29	Teachers' education level	Teaching certificate Bachelor's Master's Doctorate	Categorical	1-4	2.2
30	Teachers' L2 experience	1-5 years 6-10 years 11-15 years 16-20 years > 20 years	Categorical	1-5	2.2

Data Analysis

In this section, statistical analyses for each research question are outlined. The section starts with the preliminary analyses. Then, the following is described for each research question: dependent and independent variables with their operationalizations and levels of measurement, hypotheses, as well as statistical analyses, assumption checks, and sample size calculations. Analysis plans for each major research question end with a summary of how the expected findings contributed to the overall interpretive argument for the inclusion of content-rich videos in L2 academic listening constructs.

Preliminary analyses. *Descriptive statistics.* Descriptive statistics for the anchor and ALC tests were given separately. Sample sizes, means, standard deviations, and 95%

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

confidence intervals for the means were reported overall as well as by mode (i.e., video-based and audio-only) and location (i.e., Mexico, USA, Russia, and Facebook).

Psychometric properties. As a preliminary stage, the psychometric properties of the anchor test and the ALC test were examined separately using both Rasch and classical analyses. The Rasch analysis was run using the Facets software (Linacre, 2017). The preliminary ALC Rasch model was based on the following two facets: 143 test-takers and 24 items' scores (i.e., 0 or 1). The anchor Rasch model was based on 143 test-takers and 12 items' scores (i.e., 0 or 1). The anchor and the ALC items were analyzed for the following: (a) item infit mean square statistic, (b) item separation reliability, (c) person separation reliability, and (d) the item-ability Wright map.

Item infit mean square statistics showed the size of randomness, or distortion, in the measurement. Values higher than one generally indicated that an item lacked predictability while values lower than one indicated redundant items. The recommended range for mean square infit value was 0.75-1.30, as suggested by McNamara (1996). Misfitting items would suggest the presence of construct-irrelevant variance (Baghaei, 2008; Messick, 1989). To further support construct validity, item separation reliability was expected to approach 0.80 and higher, which would be indicative of a sufficient item difficulty hierarchy. Person separation reliability is a Rasch equivalent of Cronbach's alpha ("Reliability and Separation of Measures," 2017). Lower values (< 0.70) may imply that the instrument cannot consistently distinguish between low and high performers. All the aforementioned analyses are listed in Table 3.17. The table also provides the recommended ranges of values for each psychometric parameter.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Item analysis parameters based on the classical test theory supplemented the Rasch analysis. The following parameters were estimated: (e) item difficulty, (f) item discrimination, (g) distractor analysis, and (h) Cronbach's alpha internal consistency reliability. The range for item difficulty was set at 0.25-0.75, guarding against extremes in item difficulties. The minimum value for item discrimination was 0.25, with an ideal of 0.30 and above (Fulcher, 2010). A distractor analysis was used to reveal ineffective distractors. A distractor was deemed ineffective if it failed to attract at least 10% of test-takers' responses (Fulcher, 2010). Cronbach's alpha indices were expected to be at least 0.70, following the rule of thumb for classical reliability analysis (Nunnally & Bernstein, 1994). The recommended values for item difficulty, item discrimination, distractor analysis, and Cronbach's alpha are also listed in Table 3.17.

Table 3.17

Expected Psychometric Properties for the Anchor and ALC Tests

Approach	Analysis	Expectation/Range
Item response theory (Rasch)	(a) Item mean square infit statistics	0.75-1.30
	(b) Item separation reliability	≥ 0.80
	(c) Person separation reliability	≥ 0.70
	(d) Item-ability map	<ul style="list-style-type: none"> • person abilities and items difficulties are well-matched • there is no considerable gaps between items on the item difficulty continuum
Classical test theory	(e) Item difficulty	0.25-0.85
	(f) Item discrimination	≥ 0.25
	(g) Item distractor analysis	Each distractor attracts at least 5% of test-takers' responses
	(h) Internal consistency reliability	≥ 0.70

Determination of group equivalence. To confirm group equivalence, the audio and video groups' anchor scores were compared using an independent *t*-test. The data for

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

the *t*-test were checked for independence of observations, significant outliers, normality of score distributions in each group, and homogeneity of error variances. The normality assumption was checked with normal Q-Q plots, skewness and kurtosis statistics, and Shapiro-Wilk tests. Homogeneity of variance was examined with Levene's test.

Operationalization of test-takers' proficiency. As part of the preliminary analysis, proficiency was operationalized based on person ability logits generated by Rasch analysis for the anchor test. It was operationalized using the person ability logit mean as a cut-off point, which represents an average ability (McNamara, 1996). Test-takers with higher-than-cut-off logit values were allocated to the higher proficiency group (above average), with the rest assigned the lower proficiency category (below average).

Research question 1. Do content-rich videos affect L2 academic listening comprehension difficulty? This question was subdivided into three subquestions, each of which is described below in terms of hypotheses, analyses, and assumptions. This section concludes with a brief summary of how the expected findings would advance the argument for including content-rich visuals in the L2 academic listening construct.

Research question 1.1. Is academic listening comprehension difficulty at the test level affected by delivery mode, listening proficiency, item video-dependence, and item type? Content-related visual information is generally believed to decrease listening comprehension difficulty (Rost, 2016). Lower-level learners' comprehension may be adversely affected by the presence of visuals (Mayer, 2005; Paivio, 1991; 2006). This hypothesis, however, likely depends on how low test-takers' proficiency level is. For example, it may not hold true for intermediate learners but could work for beginners due to the latter having significantly lower language processing capacity. In any scenario,

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

item difficulties were likely to be affected by the presence of content-related visual information in a test, either negatively or positively, depending on test-takers' listening proficiency levels. It was also predicted that content-related visuals would decrease difficulty of both local and global comprehension items for higher-level test-takers, with a more noticeable effect on global items. In contrast, content-related visuals might increase difficulty of both local and global comprehension items for lower-level test-takers, with a more prominent effect on global items (Becker, 2016; Hansen & Jensen, 1994; Shohamy & Inbar, 1991). For both proficiency groups, the described effects were expected only on the video-dependent group of items. No effect was expected on video-independent items, regardless of proficiency and item type.

Both classical and Rasch approaches were used for answering research question 1.1. The classical approach built on analyses of variance (ANOVA). The Rasch approach relied on running a one-parameter Rasch model. Both approaches are described below.

Classical analyses. Seven separate ANOVAs were run to answer research question 1. Because item video-dependence and item type were properties of items but listening difficulty and proficiency were properties of test-takers, it was not possible to include video-dependence and item type as factors in one omnibus ANOVA. However, it was possible to run separate ANOVAs to analyze test-takers' responses on each subset of items, as indicated in Table 3.18. The table shows the number of items and score range for each of the seven subsets. Seven respective dependent variables were operationalized by test-takers' total scores on the respective subsets of items. All the dependent variables were treated as interval between pairwise differences between adjacent score points were the same (i.e., the difference between 6 and 7 was the same as between 21 and 22).

Table 3.18

Collection of ANOVAs for Research Question 1.1

ANOVA #	DV	<i>k</i>	Range
#1	all items	24	1-24
#2	video-dependent	14	1-14
#3	video-dependent local	7	1-7
#4	video-dependent global	7	1-7
#5	video-independent items	10	1-10
#6	video-independent local	5	1-5
#7	video-independent global	5	1-5

Note: DV = dependent, or response, variable; *k* = number of items;
 Independent variables for each ANOVA: mode (i.e., audio-only vs. video-based)
 and proficiency (i.e., lower vs. higher).

Power analysis for a 2x2 ANOVA was carried out in G*Power to determine a sufficient sample size using a .05 alpha, a 90% power, and a medium effect size (partial $\eta^2 = 0.10$) (G*Power 3.1.9.3, 2014; Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007). The required sample size was 109.

The ANOVA analyses were preceded by the assumption check for independence of observations, significant outliers, normality of the dependent variable’s distribution for each combination of the groups of the independent variable, and homogeneity of error variances. Normality was checked using normal Q-Q plots, skewness and kurtosis values, and Shapiro-Wilk tests. Homogeneity of variance was checked with Levene’s test.

Though normality is normally checked for classical analyses, ANOVA is known to be robust against violations of normality in terms of type I errors (Blanca, Alarcon, Arnau, Bono, & Bendayan, 2017; Glass, Peckham, & Sanders, 1972). Therefore, the data were not disqualified from ANOVAs if normality was violated.

The interpretation of the ANOVA analyses started with the mode-proficiency interaction as a higher-order term. If the interaction was significant, subsequent *post-hoc* Bonferroni-corrected pairwise comparisons of simple effects were run. A simple effect is

the effect of one level of a first independent variable on one level of a second independent variable (e.g., the effect of the video delivery mode on the lower proficiency level). In case of an insignificant interaction, only the main effects for delivery mode were interpreted by comparing the mean values for the audio-only and the video-based conditions.

Note that ANOVAs were run in place of originally-projected multiple regression analyses. The assumption of no multicollinearity was not met for regression because the interaction term had to be included in the model. Higher-order terms in regression models are a well-known threat to the multicollinearity assumption because they are a product of and, thus, are highly related to the original predictors (Tate, 1984). Therefore, ANOVAs were preferred to regression analyses.

Rasch analyses. A multi-faceted Rasch analysis (MFRM) was used to supplement the classical statistical analyses. Rasch analysis has several advantages over classical-test-theory-based analysis. It is said to be more linear as it relies on truly continuous data and is less test-dependent (Wright, 1992). It also allowed for deeper interpretations of video effects by running bias/interaction analyses at the item level.

The Rasch model was based on the following six facets: test-takers (i.e., 1-120), delivery mode (i.e., audio-only vs video-based), item video-dependence (i.e., video-dependent vs video-independent), item type (i.e., local vs global), test-taker proficiency (i.e., lower vs higher), and 24 items' scores (i.e., 0 or 1). The test-taker facet was non-centered. Conventionally, the agents of measurement (i.e., items, tasks, or judges) establish the origin, or frame of reference, and, therefore, are centered. The objects of measurement (i.e., test-takers) are positioned relative to the origin, and, therefore, are

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

non-centered, or “floating.” A Rasch model must have one non-centered facet to ensure that estimates are sufficiently constrained and non-ambiguous (Linacre, 2012b).

To use the Rasch model, the following assumptions had to be met: (1) unidimensionality, requiring all the items to measure the same underlying variable, and (2) local item independence, requiring every item to be independent of the others in a test. Though data from testlet-based measurements may not completely satisfy the assumption of local independence (So, 2010), the Rasch analysis can be used in such situations as long as items do not cue one another.

The dependent variable was listening comprehension. It was operationalized by item difficulty logit values collectively (i.e., item logits averaged for combinations of levels of the independent variables). Item difficulty logits were on a continuous scale of -2.06 to 4.44. Delivery mode was a categorical independent variable with two values, namely audio-only and video-based. Proficiency level was determined using the anchor test. It had two dichotomous values, lower and higher, as described in the Determination of test-takers’ proficiency section. Item video-dependence was a yes-or-no property of each individual item. Accordingly, it had two values, namely video-dependent and video-independent, reflecting the video-dependence grouping decisions for each item (see Measures). Finally, item type had two values, namely global and local, according to the ALC test specifications (see Appendix D). Note that while the dependent-vs-independent variable distinction is rarely used in Rasch analysis, it was employed in this study in order to facilitate the reader’s understanding.

To detect the effects of the independent variables on listening difficulty, a measurement report for the mode facet was examined first. A significant separation index

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

($p \leq .05$) indicated a difference between the overall mode difficulties. To determine the directionality of this difference, the difficulty logits for the audio-only and the video-based modes were compared.

To further gauge the effect of mode in relation to proficiency, item video-dependence, and item type, a number of corresponding Rasch bias/interaction analyses were run, including interactions between delivery mode and (a) proficiency, (b) video-dependence, (c) video-dependence and proficiency, (d) video-dependence and item type, and (e) video-dependence, proficiency, and item type. In these interactions, the facets represented the corresponding independent variables. The dependent variable was represented by item difficulty logits taken collectively for each combination of the levels of independent variables in an interaction. A sample Rasch specification file for Facets is found in Appendix K.

The sample size for Rasch models is conventionally determined by item calibration stability (Linacre, 1994). To achieve item calibration stability within 0.5 logits based on a 95% confidence interval, 100 observations were needed. However, this rule may not accurately apply in contexts of measuring group differences as opposed to analyzing item attributes. Research shows that sample sizes for detecting group differences using Rasch models should be 25-35% greater than sample sizes for classical analyses, assuming the same statistical power (Sébille, Blanchin, Guillemin, Falissard, & Hardouin, 2014). Based on the required sample size of 109 for the classical ANOVA analyses, the Rasch analysis required about 137-147 participants to achieve a 90% power.

Research question 1.2. Is academic listening comprehension difficulty at the item level affected by delivery mode, listening proficiency, item video-dependence, and item

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

type? For this question, there were 24 dependent variables, matching the number of items in the ALC test and operationalized as 24 item difficulty logits. Each was on a continuous scale from -1.69 to 0.99. Delivery mode, listening proficiency, item video-dependence, and item type were operationalized identically to research question 1.1 (Rasch analyses). The Rasch model was based on the same six facets as in research question 1: test-takers, delivery mode, listening proficiency, item video-dependence, item type, and item scores. Rasch bias/interaction analyses centered on each individual item difficulty rather than treating item logits collectively. Rasch bias/interactions were run between delivery mode and listening proficiency for each item. A mode effect for an item was considered present if it reached statistical significance at the .05 *alpha* level.

The associations of these interactions with item video-dependence and item type were determined descriptively, using items' specifications. Each item's difficulty was compared by mode within each combination of proficiency and item video-dependence, while also considering item type. Conclusions about corresponding trends were made.

It was hypothesized that most video-dependent items would be easier in the video-based mode than in the audio-only mode for higher-level test-takers, with a more conspicuous effect on global items. For lower-level test-takers, most video-dependent items were expected to be harder in the video-based mode than in the audio-based mode, with a more prominent effect on global items (Becker, 2016; Hansen & Jensen, 1994; Mayer, 2005; Paivio, 1991; 2006; Rost, 2016; Shohamy & Inbar, 1991). The expectations for the lower-level test-takers rested on the assumption that test-takers' proficiency levels were low enough to hinder successful dual-channel processing. No effect of delivery mode was predicted for video-independent items, regardless of proficiency and item type.

Research question 1.3. Is academic listening difficulty related to viewing behavior and listening proficiency? Viewing behavior is assumed to be a construct-relevant factor (Wagner, 2007). Considering the discussions above, it might be expected that lower-level test-takers' viewing behavior adversely related to testlet scores while higher-level test-takers' viewing behavior positively related to testlet scores. To answer research question 1.3, multiple regression was set to be used initially. However, the viewing behavior and ALC test scores were not linearly related, violating the fundamental assumption for linear regression. Instead, three Spearman's Rank-Order correlation analyses were used, one for the lower-proficiency group, one for the higher-proficiency group, and one overall. The only assumption for Spearman's correlation is having two ordinal, interval, or ratio variables. Viewing behavior composite ratings and ALC scores were assumed to be interval.

Viewing behavior was operationalized as self-reported viewing behavior ratings on question A of the test-takers' questionnaire (video-based version; see Appendix I) provided on a scale from 1 to 5 after each of the four testlets in the video-based version of the ALC test. These four ratings were summed across the four testlets for each test-taker. Therefore, viewing behavior was a continuous variable varying from 1 to 20. The summing across the testlets was justified since each testlet-based item was designed to measure exactly the same construct. Test difficulty was operationalized as a sum of test-takers' total scores on the four testlets. Therefore, this variable was identical to variable 1 in the ANOVA analyses for research question 1 (labeled as DV #1a in Table 3.15 on p. 112). Listening proficiency was operationalized in the same way as for research question

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

1 (see IV #2 in Table 3.16 on p. 113). Testlet difficulty was then correlated with viewing behavior by proficiency and overall.

Power analysis for the correlation analysis was conducted in G*Power to determine a required sample size using an alpha of .05, a power of 0.90, and a medium effect size ($r^2 = 0.30$). The required sample size was 30.

Summary. The first research question sought to generate discriminant evidence for item difficulties in the video-based mode and the audio-only mode. The video-based test version represented a sufficient, theory-informed, visually-rich L2 academic construct. The audio-only test version represented a deficient, visual-free construct. Divergent test and item difficulties by mode were expected as evidence supporting the explanation inference that links comprehension difficulty to construct-relevant factors, assuming the results were in line with the aforesaid proficiency-related theoretical anticipations. The analysis of test-takers' viewing behavior was projected to generate both divergent and convergent evidence at lower- and higher-level proficiency respectively, providing additional backing for the explanation inference assumption.

Research question 2. Do stakeholders' perceptions lend support for using content-rich videos in the L2 academic listening assessment construct? This research question was subdivided into research questions 2.1 on test-takers' perceptions and 2.2 on teachers' perceptions. They are described below with regards to hypotheses, analyses, and assumption checks. A brief summary concludes this section.

Research question 2.1. Do delivery mode and listening proficiency affect test-takers' perceptions about listening difficulty, motivation, and authenticity, and use of content-rich videos in tests? For this question, perceptions of test-takers in the audio-only

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

group were compared to those in the video-based group. Perceived listening difficulty in the video-based mode was hypothesized to be lower than difficulty in the audio-only mode for higher-level students. Lower-level test-takers might find the content-rich video mode to be more difficult than the audio-only mode due to their limited processing capacities (Mayer, 2005; Paivio, 1991; 2006; Rost, 2016). Motivation, authenticity, and video use perceptions were expected to be more favorable in the video-based mode than in the audio-only mode, regardless of test-takers' proficiency.

There were four dependent variables in the analyses for this question: listening difficulty, motivation, authenticity, and use of videos in tests. Listening difficulty was operationalized as a sum of the four 1-to-6-scale ratings elicited after each of the four ALC testlets on question 2 of the test-takers' questionnaire (both versions; see Appendix D). It ranged from 1 to 24. Motivation was operationalized as a sum of scores on question 1 of the test-takers' questionnaire across the four ALC testlets. It ranged from 1 to 24. Authenticity was operationalized as a sum of scores on question 3 of the test-takers' questionnaire across the four ALC testlets. It ranged from 1 to 24. Use of videos in tests was a sum on items 4-6 in the test-takers' questionnaire. Its values ranged on a continuous scale from 1 to 18. There also were two independent variables. The first independent variable, delivery mode, was operationalized as either audio-only or video-based (#1a in Table 3.16 on p. 113). The second independent variable, test-takers' proficiency ranged from lower to higher (#2 in Table 3.16 on p. 113).

Four ANOVAs were run to investigate the effect of mode and proficiency on the four dependent variables. Each ANOVA analysis was preceded by checking the following assumptions: independence of observations, no significant outliers, normality,

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

and homogeneity of variance. Normality and homogeneity were checked in the same manner as for research question 1 (see the Research question 1.1 section). The interpretation techniques and required sample size ($N = 109$) were also the same.

Research question 2.2. How does teachers' background (i.e., geographical region, education, and L2 teaching experience) affect their perceptions about the effect of content-rich videos on listening difficulty, motivation, authenticity, and use of content-rich videos in tests? This question tested the pilot-informed hypothesis that, regardless of their background, L2 teachers would consider viewing content-rich visuals as decreasing listening difficulty, and increasing motivation and authenticity, and would have favorable opinions about using content-rich videos in high-stakes listening tests.

There were four dependent variables: listening difficulty, motivation, authenticity, and use of content-rich videos in tests. Listening difficulty was operationalized as a sum of the four 1-to-6-scale ratings on questions 1, 5, 12, and 15 in the teachers' questionnaire (see Appendix J). Motivation was operationalized as a sum of scores on questions 2, 6, 9, and 16. Authenticity was operationalized as a sum of scores on questions 3, 7, 10 and 13. Use of videos in tests was a total for items 4, 8, 11 and 14 in the teachers' questionnaire. All the four variables ranged on a continuous scale from 1 to 24.

There were also three independent variables: geographic region, education level, and L2 experience. Geographic region had five values, namely (1) Asia and Oceania, (2) Europe and Eurasia, (3) Caribbean, Central, and South America, (4) Africa and the Middle East, and (5) North America. Education level was also a categorical variable with four possible values: 1-Teaching certificate, 2-Bachelor's, 3-Master's and 4-Doctorate. L2 experience had five values (i.e., 1-5, 6-10, 11-15, 16-20, and > 20 years).

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

After checking the assumptions, four ANOVAs were run to detect the effects of region, education, and experience on the four dependent variables. The assumption check and interpretation were similar to research questions 1.1 and 2.1.

Power analysis for a 5x4x5 ANOVA was carried out in G*Power to determine a sufficient sample size using a .05 alpha, a 90% power, and a medium effect size (partial $\eta^2 = 0.10$). The required sample size was 172. Another important parameter for questionnaire studies is a margin of error. For a pre-set margin of 5%, a population size of about 50,000, and a 95% confidence interval, the required sample size was 382 (“CheckMarket. Sample Size Calculator,” 2017).

Summary. One of the assumptions related to the explanation inference in the argument-based validity framework states that the test construct is perceived favorably by test stakeholders (e.g., test-takers, teachers). The investigation of test-takers’ and teachers’ perceptions of content-rich videos generated evidence backing this assumption. Test-takers’ and teachers’ favorable opinions about using content-rich videos in listening tests further supported the argument.

Alpha level and effect sizes. Unadjusted alpha level of .05 was used in this study despite conducting several statistical tests. While it is conventionally recommended to reduce the alpha level by the overall number of conducted tests in a study, some authors have a different view (e.g., Anderson, 2014; Ha & Ha, 2012; Tucker, 1991). They noted that, while *post-hoc* pairwise comparisons should be alpha-adjusted, planned, or *a priori*, comparisons could be exempt from alpha adjustment. Following the convention, the alpha level in this study would be significantly reduced due to numerous statistical tests employed. While helping to avoid type I error rates (i.e., false positives), it would

significantly increase type II error rates (i.e., false negatives, or missed true discoveries). Following the recent views on alpha adjustment, this study used the alpha level of .05 for planned comparisons (including Rasch interactions) and Bonferroni-adjusted alpha rates for *post-hoc* analyses. It should be noted that the ultimate decision to bypass alpha adjustment also rested on the fact that this study was of the exploratory nature and aimed to find promising patterns regarding the use of content-rich videos in listening tests.

Effects sizes for classical analyses were estimated. For ANOVAs, partial eta squared values were calculated and interpreted following Cohen's suggestions (1988). Values near 0.01, 0.06, and 0.14 indicated small, medium, and large effects respectively. For correlation, coefficients near 0.10, 0.30, and 0.50 showed small, medium, and large effects respectively (Cohen, 1988).

Chapter 4

Results

This study aimed to find evidence supporting the inclusion of content-rich videos in second language (L2) academic listening tests. Two major pieces of evidence were sought. First, an online academic English listening comprehension (ALC) test was developed and administered to English as a second and foreign language (ESL/EFL) learners in either audio-only or video-based mode. The two modes were compared for difficulty, generating an answer for the first major research question in the study: *Do content-rich videos affect L2 academic listening comprehension difficulty?* Second, online questionnaires eliciting perceptions about the two modes were developed and administered to test stakeholders, including both ESL/EFL learners and teachers. The stakeholders' perceptions were compared by mode, generating an answer for the second major research question in the study: *Do stakeholders' perceptions lend support for using content-rich videos in the L2 academic listening assessment construct?*

The results of the statistical analyses for each research question are detailed in this chapter. Starting with the data screening strategies, this chapter then reports on psychometric properties of the measures, including the anchor test and the ALC test, and on the results of each research question.

Data Screening

This section describes techniques that were used to detect data abnormalities, the process known as data screening and recommended as a forerunner to inferential statistical analyses (Tabachnik & Fidell, 2013). The data were screened for quality,

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

accuracy, missing responses, and outliers. The results for each of these categories are presented below.

Data quality. Three specific techniques were used for data quality screening. First, details automatically collected online for each response were examined, including date submitted, time started, IP address, http referrer, country, region, city, and postal code. Coupled with self-reported data, such as names and e-mail addresses, these details helped to detect duplicates and related responses. Out of the same individual's responses to test and questionnaire items, only the earliest were included in the analysis. If the earliest response was incomplete but later responses were complete, all responses were excluded from the study. Eight responses were excluded from the study.

Second, there were three data-quality items that test-takers answered upon finishing the assessment battery. One item asked about problems with technology, including slow internet connection, glitches, and slow videos. If a participant reported a problem, he or she was asked a follow-up question about whether the problem significantly affected test performance. In case of a positive answer or no answer from the participant, his or her data were excluded. Note that some comments left for this item were not considered problems, including "I want to watch video," "no video," "speed of speakers varies," "the lectures are read so fast," "the last audio was not as loud as others," "it's too difficult." Such comments either described technology-unrelated sentiments or minor issues unlikely to affect test scores. Some of the mentioned issues were just conditions determined by delivery mode (e.g., "no videos"). The other two items asked whether the test-taker had paused the test or whether the test-taker reloaded web-pages while taking the test. In case of a positive answer for either of these items, the

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

participant's response was excluded from the analysis. Following these criteria, seventeen responses were excluded from the study.

As the third technique for data quality screening, some self-reported demographics, such as age and first language, were checked. Despite confirming their age (i.e., 18 years old) when electronically signing the informed consent before taking the test, some participants indicated a younger age for the post-test age-related demographic item. Such participants were considered minors. Their data were destroyed (5 responses), as directed by the university's IRB office. Data from participants who reported English as their first language were excluded from the study (3 responses) except for one participant for whom it was an obvious mistake (i.e., the researcher knew no native speakers were tested, and the participant's overall score for the test was very low).

In total, 177 completed responses were examined. Out of the 177 files, 34 were excluded from the study (about 19%).

Data accuracy. To screen the collected data for accuracy, two techniques were used. First, data were checked to ensure accurate coding entry. For this, minimum and maximum data values were examined for each measurement instrument, namely the ALC test, the anchor test, the test-takers' questionnaire, and the teachers' questionnaire. The corresponding minimums and maximums had expected values.

Second, overall score ranges for the academic listening comprehension (ALC) and anchor tests were examined. It was found that the respective overall scores fell within expected ranges: 3-24 for the ALC test (out of 0-24) and 0-12 for the anchor test (out of 0-12). For the test-takers' and teachers' questionnaires, the distribution of responses for

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

each question were examined. It was found that scores for each individual item were expectedly distributed within the range of 1-6.

Missing data. Next, the data were screened for missing values for each of the four measurements. One response from the ALC test was missing, which was 0.0002% of the collected ALC test data. There were no missing values in the anchor test responses, in the test-takers' questionnaire responses, and in the teachers' questionnaire responses. Because there was no discernable pattern of missing data, the missing data were considered Missing Completely at Random (MCR). The only missing value in the ALC test was treated as incorrect and, therefore, coded with a zero.

Outliers. Outliers are extreme values in the dataset that can skew statistical findings. Tabachnik and Fidel (2013) suggested standardizing raw scores and removing data points exceeding the absolute value of 3.29. Using this method, no outliers were detected for any variable related to the ALC test, the anchor test, or the test-takers' questionnaire. In contrast, a number of outliers were detected in the teachers' questionnaire data, including three outliers for teachers' difficulty perceptions, three for motivation, two for authenticity, and three for video use. These outliers were excluded from the corresponding analyses, even though this resulted in different sample sizes.

Preliminary Analysis for the Anchor Test

As part of the test-takers' assessment battery, the ALC test was administered first, followed by the anchor test. Despite this, the anchor test is described first as it steered the use of the ALC test scores. The anchor test was used to examine equivalence of the audio-only and the video-based groups and to determine proficiency levels of test-takers. Recall that all participants took the same anchor test regardless of their assignment to

treatment groups (i.e., audio-only vs. video-based). The anchor test consisted of two lectures, each followed by six multiple-choice questions. This section examines psychometric properties of the anchor test first, followed by the reports on listening proficiency and group equivalence by mode.

Psychometric properties. Informed by both classical and Rasch analyses, this section describes item-level statistics and results of the reliability analyses for the anchor test. It starts with the descriptive statistics, goes through item and reliability analyses, and ends with a brief summary.

Descriptive statistics. Measures of central tendency (i.e., means and standard deviations) along with sample sizes and confidence intervals are given in Table 4.1 for each delivery mode and location of participants. The column totals show that the audio-only mode generated somewhat higher anchor scores in all locations except for Russia, where it was slightly harder. However, largely overlapping confidence intervals suggest that the mode-based differences within each location were not considerable. The row totals show that the two modes' averages were similar. The column totals show that Facebook participants were the most proficient, followed by Mexico, Russia; the USA-based participants were the least proficient. According to the grand total mean, test-takers got about half of the 12 anchor items right.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Table 4.1

Descriptive Statistics for Anchor Test Scores

Delivery mode		Participants' location				Total (k=12)
		Mexico	USA	Russia	Facebook	
Audio-only	<i>n</i>	35	9	20	11	75
	<i>M</i>	7.09	5.00	5.95	8.46	6.73
	<i>SD</i>	2.32	2.06	2.95	1.97	2.59
	CI	[6.27; 7.90]	[3.40; 6.60]	[4.87; 7.03]	[7.70; 9.91]	[6.13; 7.33]
Video-based	<i>n</i>	39	9	18	2	68
	<i>M</i>	6.80	4.78	6.56	7.50	6.49
	<i>SD</i>	2.39	1.86	2.68	3.54	2.48
	CI	[6.02; 7.57]	[3.17; 6.38]	[5.42; 7.69]	[4.10; 10.90]	[5.89; 7.08]
Total	<i>n</i>	74	18	38	13	143
	<i>M</i>	6.93	4.89	6.24	8.31	6.62
	<i>SD</i>	2.34	1.91	2.80	2.10	2.53
	CI	[6.39; 7.49]	[3.94; 5.84]	[5.32; 7.16]	[7.04; 9.57]	[6.20; 7.03]

Note: *n* = sample size; *M* = mean; *SD* = standard deviation; CI = confidence interval; *k* = number of items

Item performance statistics. To ensure that the anchor test was appropriate for making norm-referenced decisions, psychometric qualities of its items were examined. Table 4.2 displays the following information for each of the 12 anchor test items: Rasch item difficulty logits, Rasch infit mean square statistics, classical item difficulty, classical item discrimination, and distractor choice proportion values. The anchor difficulty logits ranged from -1.44 to 2.59. This may show that the anchor items had strong potential to differentiate between proficiency levels. All the items had infit mean square values within the recommended range of 0.75 to 1.3. This indicates that the anchor test items worked well with one another, forming a stable unidimensional measure.

Means and standard deviations in Table 4.2 show that the two anchor testlets were very similar in terms of difficulty, with both testlets being at a medium level (i.e., close to the zero difficulty value). On average, both Cybersecurity and Language had acceptable discrimination indices of no less than 0.25, *M* = 0.26, *SD* = 0.11 and *M* = 0.33, *SD* =

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

0.11 respectively (Fulcher, 2010). Some distractors attracted less than 10% of test-takers' responses, with three of them approaching the critical zero value (items 1, 2, and 7).

According to both Rasch and classical analyses, one item was especially difficult for test-takers (item 5) while two items were easier than recommended (items 2 and 6). The anchor item Rasch logit range was -1.44 to 2.59 (interpretation: the greater, the harder) and classical difficulty range was as wide as 0.14-0.79 (interpretation: the smaller, the harder), supporting the test's suitability to estimate test-takers' proficiency.

Table 4.2

Anchor Test Items' Properties

Testlet	Item	Rasch analysis		Classical Analysis		Distractor Analysis			
		Difficulty	Infit MS	Difficulty	Discrimination	A	B	C	D
Cybersecurity	1	-0.87	0.81	.71	.47	.21	.01*	.08*	.71
	2	-1.44	1.02	.79*	.27	.03*	.08*	.79	.10
	3	0.25	1.02	.50	.29	.15	.50	.23	.11
	4	1.03	1.12	.36	.15	.43	.11	.36	.10
	5	2.59	1.06	.14*	.20	.22	.52	.11	.14
	6	-1.38	1.16	.78*	.17	.78	.06*	.05*	.10
	<i>M</i>	0.03	1.03	.55	.26				
	<i>SD</i>	1.59	0.12	.26	.11				
Language	7	-0.78	0.85	.69	.43	.20	.07*	.69	.03*
	8	1.15	1.01	.34	.30	.48	.08	.34	.11
	9	-0.30	1.05	.61	.26	.22	.09*	.61	.08*
	10	-0.34	0.87	.62	.42	.17	.06	.16	.62
	11	0.10	0.89	.53	.41	.11	.53	.08*	.28
	12	-0.01	1.15	.55	.17	.06*	.08*	.55	.31
	<i>M</i>	-0.03	0.97	.56	.33				
	<i>SD</i>	0.65	0.12	.12	.11				
Total	<i>M</i>	0.00	1.00	.55	.30				
	<i>SD</i>	1.16	0.12	.19	.11				

Note: *M* = mean; *SD* = standard deviation; * = outside the recommended range

Next, the person-item map, or the “Wright map”, was examined (McNamara, 1996; Wilson, 2011; see Figure 4.1). The map depicts Rasch person and item logits along the same continuum. Negative person logits indicate lower ability while positive logits

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

show higher ability. Negative item logits show easier items, and positive logits show harder items.

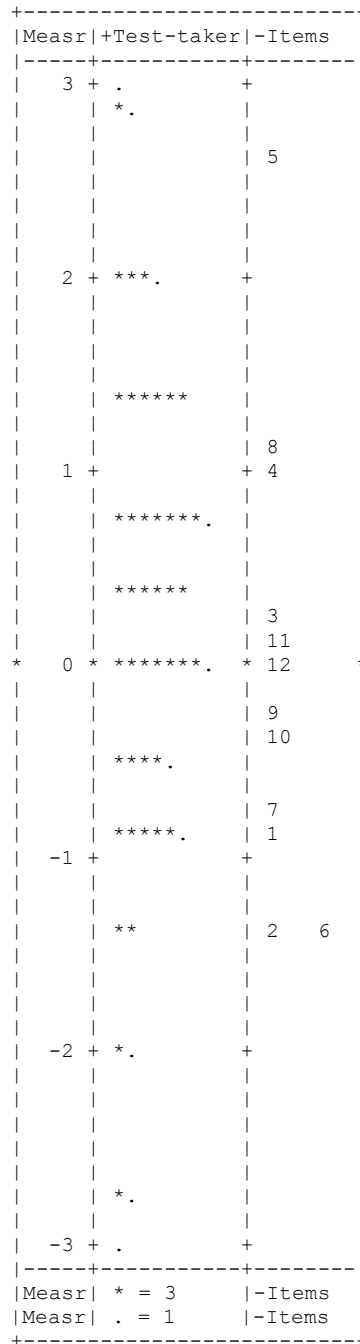


Figure 4.1. The Wright map for the anchor test.

Figure 4.1 suggests that the anchor items were well-matched with test-takers' listening abilities on the whole. Item difficulties were spread out and mapped to both

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

negative and positive person abilities. However, the test-takers’ abilities varied within a wider range than item logits did. Specifically, we see no items aligned with nine lower-level test-takers, showing that the test may need a few more low-difficulty items. There are also gaps in the higher end of the items continuum, where only one item was matched with 33 test-takers. This implies that whereas the item difficulties ranged widely, some test-takers’ abilities did not have items of matching difficulties. Understandably, it was a challenge for a test of only 12 items to closely tap on the variety of test-takers’ abilities.

Reliability analyses. The Rasch person reliability was 0.66 while the Rasch item reliability was 0.96. Coupled with the strata value of 2.18 (see Table 4.3), the first index shows that the anchor test could reliably distinguish between at least two proficiency levels of test-takers (Linacre, 2012a; 2013). The second index indicates an effective hierarchy of the anchor test items in terms of difficulty (strata = 7.55). The classical reliability analysis for the anchor test was done by computing a Cronbach’s alpha internal consistency reliability value. It was 0.65.

Table 4.3

Anchor Test: Summary Statistics for Rasch Test-Takers and Items Facets

Facet	<i>N</i> / <i>k</i>	<i>M</i>	<i>SD</i>	Mean Infit MS	Mean Infit Z	Reliability of separation	Strata	Chi-square (fixed)	<i>df</i>	<i>p</i>
Test-takers	143	0.27	1.32	0.99	0.00	.66	2.18	296.3	142	< .01*
Items	12	0.00	1.16	1.00	0.00	.97	7.55	264.7	11	< .01*

Note: *n* = sample size; *k* = number of items; *M* = mean; *SD* = standard deviation; MS = Mean Square; Z = standardized value; *df* = degrees of freedom; * significant at the .05 alpha level; The test-taker facet was centered. The Items facet was non-centered.

Summary. The psychometric properties of the anchor test were largely within the recommended ranges. Individual items’ fit into the measured construct was good. Difficulty and discrimination values showed items’ adequate capacity to distinguish among test-takers’ listening abilities. The anchor test demonstrated a facility to reliably differentiate between at least two proficiency levels. All this renders the anchor test

suitable as a measure of test-takers' listening proficiency.

Determination of group equivalence. Prior to answering research question 1, it was necessary to check the assumption that the video-based and the audio-only groups of test-takers had comparable proficiency. It was done by running an independent-samples *t*-test on anchor total scores. The assumptions check for the *t*-test included independence of observations, no significant outliers, normality of score distributions in each group, and homogeneity of variances. Independence of observations was warranted since the two mode-based groups were non-overlapping and unrelated. To check for outliers, the anchor test scores were converted into *z* scores. There were no scores outside the absolute value of 3.29, indicating the absence of outliers.

The normality assumption was examined with normal Q-Q plots of *z* scores, skewness/kurtosis statistics, and Shapiro-Wilk tests for the audio-only and the video-based groups separately. Evidence based on these sources revealed no serious deviations from normality. The homogeneity-of-variance assumption was also met, as indicated by Levene's test, $F = 0.16, p = .69 > .05$.

The *t*-test returned no differences between the anchor total scores of the audio-only group ($n = 75, M = 6.73, SD = 2.59$) and the video-based group ($n = 68, M = 6.49, SD = 2.48$), $t(141) = 0.58, p = .56 > .05, 95\% \text{ CI } [-0.59, 1.09]$. Thus, it was assumed that the two groups were not different in terms of academic listening proficiency.

Operationalization of test-takers' proficiency. In order to include proficiency in subsequent analyses, it was operationalized based on person ability logits obtained from the Rasch analysis on the anchor test ($N = 143$). Because the anchor test had the capacity to distinguish between two proficiency levels, two categories were used to operationalize

proficiency, namely lower and higher. To assign test-takers either the lower or the higher category, the person ability logit mean was used as a cut-off point ($M = 0.27$, $SD = 1.32$ in Table 4.3). Test-takers with ability logits higher than 0.27 were grouped under the higher proficiency label, with the rest assigned the lower category. Out of 143 recruited participants, 67 were at the lower level and 76 were at the higher level.

Preliminary Analyses for the ALC Test

The academic listening comprehension (ALC) test was used for detecting the effect of mode on test-takers' listening comprehension. One group of test-takers took the ALC test in the audio-only mode while the other in the video-based mode. The performance of the two groups was then compared taking into account three other variables (i.e., test-takers' proficiency, item video-dependence, and item type). This section reports on psychometric properties of the ALC items. It also describes a specific technique used to avoid disjoint Rasch datasets.

Psychometric properties. Beginning with the descriptive statistics for the ALC test, item properties of the ALC test are then described using both the Rasch and classical approaches. This is followed by reliability analyses and a brief summary.

Descriptive statistics. The ALC total score was examined by mode (i.e., audio-only and video-based) and by location (i.e., Mexico-, USA-, Russia-, and Facebook-based), as shown in Table 4.4 below. The descriptive statistics include sample sizes, means, standard deviations, and confidence intervals.

According to Table 4.4, the audio-only ALC test was slightly easier than the video-based ALC test for participants in each location. As shown in row totals, the video-based version of the ALC test generated slightly higher scores than the audio-only

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

version. There is also a noticeable difference in the results across the four locations, with the USA-based test-takers being the lowest achievers, Russia-based participants in the middle, and the Mexico- and Facebook-based test-takers the highest. The grand total mean shows that test-takers were able to get about half of the 24 test items right on average ($N = 143$, $M = 12.28$, $SD = 4.12$).

Table 4.4

Descriptive Statistics for the ALC Test Score

Delivery mode	Participants' location				Total ($k=24$)	
	Mexico	USA	Russia	Facebook		
Audio-only	<i>n</i>	35	9	20	11	75
	<i>M</i>	12.54	8.33	10.45	12.91	11.53
	<i>SD</i>	3.23	3.67	3.94	2.55	3.67
	CI	[11.24; 13.85]	[5.76; 10.91]	[8.72; 12.18]	[10.58; 15.24]	[10.69; 12.38]
Video-based	<i>n</i>	39	9	18	2	68
	<i>M</i>	14.05	11.33	11.83	14.00	13.10
	<i>SD</i>	4.44	2.87	4.77	5.66	4.45
	CI	[12.82; 15.29]	[8.76; 13.91]	[10.01; 13.65]	[8.54; 19.46]	[12.03; 14.18]
Total	<i>n</i>	74	18	38	13	143
	<i>M</i>	13.34	9.83	11.11	13.08	12.28
	<i>SD</i>	3.96	3.55	4.35	2.487	4.12
	CI	[12.42; 14.26]	[8.07; 11.60]	[9.68; 12.53]	[11.34; 14.81]	[11.60; 12.96]

Note: n = sample size; M = mean; SD = standard deviation; CI = confidence interval; k = number of items

Item performance statistics. This section examines the suitability of the ALC test for norm-referenced interpretations. It describes how closely the ALC test items met the pre-determined criteria for both Rasch statistics (i.e., item logit difficulties and infit mean square values) and classical statistics (i.e., item difficulties, item discrimination, and distractor analysis). To obtain the Rasch infit mean square values, a one-parameter Rasch model was run. It was set on the following two facets: Test-takers (1-143) and items (1-24), with test-takers being a non-centered, or floating, facet.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Table 4.5 shows values for the Rasch and classical parameters for each ALC test item. The item difficulty logits ranged from -1.68 to 1.04. Infit item mean square values ranged from 0.86 to 1.14, which is within the norm of 0.75-1.30. As no items were misfitting, each item had a balance between predictability and variation of responses. This supports the ALC test as a predictable measure of academic listening comprehension. It also upholds the assumption of unidimensionality.

Table 4.5

Item-Level Psychometric Properties of the ALC Test Items

Testlet	Item	Rasch analysis		Classical Analysis		Distractor Analysis			
		Difficulty	Infit MS	Difficulty	Discrimination	A	B	C	D
Homeostasis	1	0.64	0.89	.38	.40	.21	.38	.16	.25
	2	-0.86	1.14	.69	.03*	.20	.05*	.69	.06
	3	0.31	0.96	.45	.31	.45	.17	.13	.25
	4	0.21	1.00	.47	.26	.47	.10	.12	.31
	5	-0.02	1.02	.52	.22	.08*	.52	.20	.21
	6	1.04	1.11	.30	.09*	.06*	.13	.52**	.30
Food Tax	7	-0.15	0.99	.55	.26	.14	.15	.16	.55
	8	-1.25	0.90	.76*	.34	.05*	.76	.16	.03*
	9	0.89	1.03	.33	.21	.17	.09*	.41**	.33
	10	0.92	1.06	.32	.19	.02*	.32	.50**	.16
	11	0.02	0.97	.51	.29	.51	.15	.14	.20
	12	0.02	1.04	.51	.21	.09*	.30	.10	.51
Compassion	13	-0.02	1.07	.52	.17	.52	.25	.10	.13
	14	0.21	1.05	.47	.19	.20	.47	.13	.20
	15	-0.44	0.86	.61	.42	.12	.61	.16	.11
	16	-0.08	1.08	.53	.15	.13	.18	.53	.15
	17	-0.08	0.93	.53	.35	.27	.12	.08*	.53
	18	-1.68	0.92	.83*	.30	.83	.07*	.05*	.06*
Exoplanets	19	-0.41	1.04	.60	.18	.21	.14	.05*	.60
	20	0.05	0.96	.50	.30	.51	.20	.11	.18
	21	0.64	1.03	.38	.22	.15	.38	.20	.27
	22	-0.44	0.96	.61	.29	.61	.08*	.13	.18
	23	0.21	0.98	.48	.28	.20	.09*	.48	.23
	24	0.27	1.00	.45	.25	.32	.45	.15	.08*

Note: * outside the recommended range; ** distractor attracting more responses than the key; $N = 143$

Classical item difficulty values were largely in the desired range of 0.25 to 0.75, except for two items. Items 8 and 18 were overly easy for test-takers. Thirteen of the 24

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

item discrimination indices were higher than 0.25. Nine of the 24 items had discrimination values in the range of .15-.22 (items 2, 5, 6, 9, 10, 12, 13, 14, 16, 19, 21). These nine items reached or approached the discrimination value of .20, which is considered acceptable by some scholars (e.g., Ebel & Frisbie, 1986; Kline, 1993) but lower than generally accepted. Two items seemed to have especially low discrimination indices (i.e., items 2 and 6). However, they were kept in the analysis because (a) their indices were greater than zero and, therefore, did not undermine the reliability of the test, and (b) the Rasch parameters for these items met the pre-set criteria.

Table 4.5 also shows the results of the distractor analysis for the ALC test items. Distractors with less than 10% of attracted responses were flagged as underperforming. About half of the items had one underperforming distractor. Item 18 had three underperforming distractors. In contrast, three distractors over-performed (see items 6, 9, and 10). Although some distractors had low attraction power, keys and distractors of the ALC test items functioned well in general. No critical problems, such as distractors with zero choice frequency or items with three non-functioning distractors, were discovered.

Table 4.6 shows descriptive statistics for psychometric parameters in more detail to ensure that testlets and item subsets were appropriate for norm-referenced interpretations. The Compassion testlet had the lowest difficulty while Homeostasis was the hardest. The video-dependent items were easier ($M = -0.28$, $SD = 0.65$) relative to the video-independent items ($M = 0.39$, $SD = 0.40$), which was also indicated by classical item difficulty means ($M = .57$, $SD = 0.13$ and $M = .43$, $SD = 0.08$ respectively). Global items were slightly harder within both the video-dependent and video-independents subsets. The overall mean Rasch difficulty logit value was 0.00 ($SD = 0.64$), showing that

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

the ALC items were average in difficulty. Infit statistics were around the target value of 1 both overall and across the subsets of items. Thus, there were no misfitting testlets or item subsets.

Table 4.6

ALC Item Statistics by Testlet, Video-Dependence, and Item Type

Subset of items	N	k	Rasch				Classical			
			Difficulty		Infit MS		Difficulty		Discrimination	
			M	SD	M	SD	M	SD	M	SD
Testlet										
Homeostasis	143	6	0.22	0.66	1.02	0.09	.47	0.12	.22	0.14
Food Tax	143	6	0.08	0.80	0.99	0.06	.50	0.15	.25	0.06
Compassion	143	6	-0.35	0.69	0.99	0.09	.58	0.12	.26	0.11
Exoplanets	143	6	0.05	0.42	1.00	0.03	.50	0.08	.25	0.05
Video-dependent items										
Local	143	7	-0.36	0.68	0.99	0.09	.59	0.14	.24	0.13
Global	143	7	-0.19	0.67	0.99	0.05	.55	0.13	.25	0.05
Video-independent items										
Local	143	5	0.29	0.34	1.00	0.08	.45	0.07	.25	0.10
Global	143	5	0.49	0.47	1.01	0.07	.41	0.10	.24	0.10
Total	143	24	0.00	0.64	1.00	0.07	.51	0.13	.25	0.08

Note: N = sample size; k = number of items; M = mean; SD = standard deviation; MS = mean squared values

Mean classical item difficulties across testlets and item subsets ranged from .41 to .59, with the grand mean of .51 ($SD = 0.13$). This indicates that item subsets had average difficulty. Mean item discrimination values across item groups ranged from .22 to .26. Although somewhat low, they was approaching the norm. The discrimination grand mean of .25 ($SD = 0.08$) shows that the ALC test items were able to discriminate among low and high achievers.

Next, the Wright map was examined. Figure 4.2 shows that, for the most part, test-takers' abilities matched item difficulties. However, the item difficulty range was slightly narrower than the person ability range. The abilities of 18 (out of 143) test-takers

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

surpassed the levels that the items were capable of targeting, suggesting that the ALC test was somewhat easy for these test-takers.

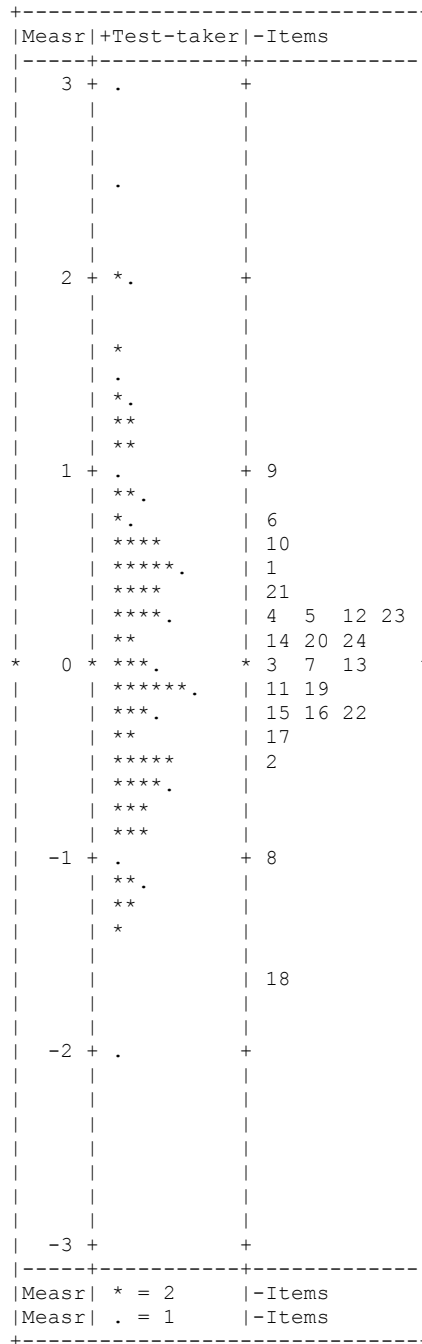


Figure 4.2. The Wright map for the ALC test items.

Reliability analyses. Rasch summary and reliability statistics for test-takers'

(person) abilities and item difficulty logits for the ALC test are shown in Table 4.7. The

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

person separation reliability was .71 (> .70). As a Rasch equivalent of Cronbach's alpha, it shows that the ALC test was able to detect 70% of the "true" variance in test-takers' scores and 30% of "noise." This parameter also shows that there were more than one distinct reproducible range in test-takers' test performance. The person reliability of no lower than .60-.70 is normally recommended as a minimum (Fisher, 2008).

Table 4.7

ALC Test: Summary Statistics for Rasch Test-Takers and Items Facets

Facet	<i>N/k</i>	<i>M</i>	<i>SD</i>	Mean Infit MS	Mean Infit Z	Reliability of separation	Strata	<i>Chi-square</i> (fixed)	<i>df</i>	<i>p</i>
Test-takers	143	0.09	0.47	1.00	0.00	.71	1.57	369.6	142	< .01*
Items	24	0.00	0.19	1.00	0.00	.92	3.31	233.1	23	< .01*

Note: *N* = sample size; *k* = number of items; *M* = mean; *SD* = standard deviation; MS = mean squared values; Z = standardized value; * = significant at the .05 alpha level

The item reliability was .92 (> .70), indicating a stable reproducible hierarchy of items. The items' strata value of 3.31 shows that the items could suit three statistically distinct proficiency levels. However, the test-takers' strata of 1.57 indicates that the hierarchy of test-takers' abilities had no more than one and a half distinct levels.

Classical reliability analyses were conducted by estimating internal consistency Cronbach's alpha coefficients. Since this study compared the performance on ALC items by delivery mode, it was most relevant to estimate internal consistency coefficients for the audio-only (.61) and the video-based modes (.75) separately. Based on the commonly recommended cut-off value of 0.70, the audio-only version had a low reliability value. The video-based version was adequate. This can be explained by the fact that many of the audio-only test items were inherently tied to (cued by) videos. In other words, they were created to go with the videos. In the absence of the videos, the items may have performed less consistently, lowering the Cronbach's alpha.

Summary. Overall, the ALC test had acceptable psychometric properties. The ALC test items (a) fit it well in the measured construct, (b) had largely acceptable difficulty and discrimination values, and (c) generated adequate reliability by delivery mode. This supports the validity of the ALC test and its suitability for norm-referenced interpretations.

Data subset connection for Rasch analysis. Recall that Rasch analysis was to be used to detect differences in difficulty between the audio-only and video-based modes of the ALC test (i.e., research question 1). The Rasch analysis for this purpose was based on a different sample configuration compared to the Rasch analysis for psychometric purposes. As a result of the first-phase recruitment of participants, the audio-only ALC test version was taken by 75 learners and the video-based version by 68 learners (143 in total). These subsamples of 75 and 68 learners did not overlap. Because of no overlap, such a configuration would render disjointed subsets in a Rasch model. While not relevant for psychometric purposes, it was desirable to connect the subsets for the mainstream analyses of mode effects.

To connect the subsets in the Rasch analysis, there had to be test-takers who would take both the audio-only and the video-based versions of the ALC test, thereby linking the mode-related data subsets. Moreover, these test-takers had to represent the two proficiency levels in order to link the proficiency-related data subsets. To achieve this, 24 out of the 143 first-phase ALC test-takers were recruited again with the purpose of taking the test in the opposite delivery mode. For example, if a test-taker took the audio-only version initially, he or she (hereafter, a repeater) was invited to take the video-

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

based version. This test-taker's results were then coded and entered into the Rasch model. Overall, 24 repeaters were recruited, which constituted nearly 17% of the initial sample.

Another requirement was that the repeaters had to represent different proficiency levels. Therefore, the recruitment of repeaters was approached strategically, as follows. Eleven of repeaters took the video-based version initially. Out of these 11 test-takers, five were in the higher proficiency group and the other six were in the lower proficiency group. The other thirteen repeaters took the audio-only version initially. Out of these 13, seven were in the higher proficiency group and the remaining six were in the lower-proficiency group. To offset the effects of memory and familiarity, the times of the two test administrations were at least three weeks apart for each test-taker. Thus, the Rasch analysis for research question 1 was based on the sample of 167 participants, 24 of whom were repeaters.

Note that other analyses in the study were based on the first-phase sample of 143 participants with no repeaters. This includes the Rasch analysis for psychometric purposes (described in the previous section), descriptive statistics, and classical analyses throughout the study. Subset connection was not required for these analyses.

To ensure that the first-phase sample and the sample with repeaters were not fundamentally different, Rasch items statistics for the two samples were compared descriptively. Table 4.8 contains the following information for Rasch difficulty logits and for Rasch infit mean square parameters: a value for the first-phase sample ($N = 143$), an analogous value for the repeaters-inclusive sample ($N = 167$), and the contrast between the two values. While values for individual items were slightly different, the parameters' mean values were identical, with the contrasts being equal to zero. This may show that

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

the two related samples yielded similar patterns of data. Thus, it was assumed that the inclusion of repeaters into Rasch analysis did not alter the nature or patterns of the data.

Table 4.8

Comparison of Rasch Properties for ALC Items between Two Samples

Testlet	Item	Difficulty logits			Infit MS		
		<i>N</i> = 143	<i>N</i> = 167	Contrast	<i>N</i> = 143	<i>N</i> = 167	Contrast
Homeostasis	1	0.64	0.49	0.15	0.89	0.86	0.03
	2	-0.86	-0.52	-0.34	1.14	1.10	0.04
	3	0.31	-0.05	0.36	0.96	0.97	-0.01
	4	0.21	0.25	-0.04	1.00	1.01	-0.01
	5	-0.02	0.20	-0.22	1.02	1.00	0.02
	6	1.04	0.81	0.23	1.11	1.08	0.03
Food Tax	7	-0.15	-0.05	-0.10	0.99	1.01	-0.02
	8	-1.25	-1.02	-0.23	0.90	0.92	-0.02
	9	0.89	0.99	-0.10	1.03	1.02	0.01
	10	0.92	0.65	0.27	1.06	1.06	0.00
	11	0.02	-0.10	0.12	0.97	0.96	0.01
	12	0.02	0.20	-0.18	1.04	0.99	0.05
Compassion	13	-0.02	-0.02	0.00	1.07	1.08	-0.01
	14	0.21	0.15	0.06	1.05	1.07	-0.02
	15	-0.44	-0.28	-0.16	0.86	0.88	-0.02
	16	-0.08	-0.21	0.13	1.08	1.09	-0.01
	17	-0.08	-0.38	0.30	0.93	0.96	-0.03
	18	-1.68	-1.68	0.00	0.92	0.92	0.00
Exoplanets	19	-0.41	-0.13	-0.28	1.04	1.07	-0.03
	20	0.05	0.17	-0.12	0.96	0.95	0.01
	21	0.64	0.43	0.21	1.03	1.01	0.02
	22	-0.44	-0.22	-0.22	0.96	0.99	-0.03
	23	0.21	0.23	-0.02	0.98	0.98	0.00
	24	0.27	0.07	0.20	1.00	1.01	-0.01
<i>M</i>		0.00	0.00	0.00	1.00	1.00	0.00
<i>SD</i>		0.64	0.56	0.20	0.07	0.07	0.02

Note: *N* = sample size; *M* = mean; *SD* = standard deviation; MS = mean square

Results for Research Question 1

Research question 1 asked about the effects of delivery mode on test-takers’ listening comprehension. It was subdivided into three subquestions, namely research questions 1.1, 1.2, and 1.3. The results of statistical analyses for each of the three

subquestions are presented in this section. Both classical and Rasch analyses were applied to research question 1.1. They are described separately.

Research question 1.1: Classical Analysis. Research question 1.1 asked: Is academic listening comprehension difficulty at the test level affected by delivery mode, listening proficiency, item video-dependence, and/or item type? To uncover the individual and joint effects of mode, proficiency, video-dependence, and item type, seven separate ANOVAs were run. The dependent variables for the seven ANOVAs are provided in Table 4.9, along with number of items and ranges. The independent variables for each ANOVA were delivery mode (i.e., audio-only vs. video-based) and proficiency (i.e., lower vs. higher). Note that these classical analyses were based on the first-phase no-repeaters sample ($n = 143$).

Table 4.9

Collection of ANOVAs for Research Question 1.1

ANOVA #	DV	k	Range
#1	all items	24	1-24
#2	video-dependent	14	1-14
#3	video-dependent local	7	1-7
#4	video-dependent global	7	1-7
#5	video-independent items	10	1-10
#6	video-independent local	5	1-5
#7	video-independent global	5	1-5

Note: DV = dependent variable; k = number of items; Independent variables for #1-7: mode (audio-only vs. video-based) and proficiency (lower vs. higher)

This section explains how assumptions for the ANOVAs were checked first. It describes the results for each ANOVA next. The section ends with a summary.

Assumption check. Prior to running each ANOVA, the following assumptions were checked: independence of observations, no significant outliers, normality of the dependent variable’s distribution for each combination of the groups of the independent variables, and homogeneity of variance for combinations of the groups of the

independent variables. Independence of observations was warranted by the design of the study. No significant outliers ($z \geq 3.29$) were found.

Normality was checked for each of the combinations of the variables using normal Q-Q plots, skewness and kurtosis values, and Shapiro-Wilk test results. For ANOVA #1 (all ALC items) and ANOVA #2 (video-dependent items), the assumption of normality was met as none of the combinations had a significant Shapiro-Wilk statistic or critically deviating values of skewness and kurtosis. The remaining ANOVAs did not meet the normality assumption. Equality of error variances was checked using Levene's test. The seven Levene's tests indicated no violations of the homogeneity assumption for any ANOVA. Since ANOVA is robust against violations of normality and the homogeneity of variance was supported, the seven ANOVAs could be run.

ANOVA #1 on all items. The video-based group scored somewhat higher than the audio-only group overall and within both of the proficiency categories, as shown in Table 4.10. The table gives descriptive statistics for each combination of delivery mode and listening proficiency as well as overall.

The overall-test ANOVA showed no significant interaction between delivery mode and proficiency, as reflected in Table 4.11. The main effect of delivery mode was significant. The video-based mode was easier than the audio-only mode (see Table 4.10, last column), with a small effect size ($\eta^2 = .05$). The significant main effect of proficiency indicates that higher test-takers outperformed lower test-takers, as expected.

The results of the first ANOVA did not support the hypothesis that the video-based mode would be easier for the higher-proficiency test-takers but harder for the lower-proficiency test-takers, as no interaction was found. The effects of mode were

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

expected to cancel out, yielding no overall effect of mode on the ALC test. This hypothesis was not supported either. Instead, the video-based mode resulted in slightly higher scores regardless of proficiency.

Table 4.10

Descriptive Statistics for 24 ALC Test Items by Mode and Proficiency

Delivery mode		Proficiency		Total ($k=24$)
		Lower	Higher	
Audio-only	<i>n</i>	33	42	75
	<i>M</i>	9.73	12.95	11.53
	<i>SD</i>	3.41	3.24	3.67
	CI	[8.52; 10.94]	[11.94; 13.96]	[10.69; 12.38]
Video-based	<i>n</i>	34	34	68
	<i>M</i>	11.44	14.76	13.10
	<i>SD</i>	4.00	4.31	4.45
	CI	[10.05; 12.84]	[13.26; 16.27]	[12.03; 14.18]
Total	<i>n</i>	67	76	143
	<i>M</i>	10.60	13.76	12.28
	<i>SD</i>	3.79	3.84	4.12
	CI	[9.67; 11.52]	[12.89; 14.64]	[11.60; 12.96]

Note: *n* = sample size; *M* = mean; *SD* = standard deviation; CI = confidence interval; *k* = number of items

Table 4.11

Results of Two-Way Factorial ANOVA for RQ 1.1: Overall ALC Scores

Source	<i>df</i>	<i>F</i>	<i>p</i>	η^2
Delivery mode	1	7.88	.006*	.05
Proficiency	1	27.17	< .001*	.16
Mode*Proficiency	1	0.01	.938	< .01
Error	139			

Note: *df* = degrees of freedom; *F* = F-statistic; * = significant at α of .05; η^2 = partial eta squared; *N* = 143

ANOVA #2 on video-dependent items. Fourteen video-dependent items were easier with videos than without across both proficiency categories and overall, as reflected in the first two rows in Table 4.12. Column totals show that higher-level test-takers outperformed lower-level test-takers on video-independent items.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Table 4.12

Descriptive Statistics for 14 Video-Dependent Items

Delivery mode		Proficiency		Total (<i>k</i> =14)
		Lower	Higher	
Audio-only	<i>n</i>	33	42	75
	<i>M</i>	6.09	8.14	7.24
	<i>SD</i>	2.43	2.13	2.47
	CI	[5.23; 6.95]	[7.48; 8.81]	[6.67; 7.81]
Video-based	<i>n</i>	34	34	68
	<i>M</i>	8.00	9.53	8.76
	<i>SD</i>	2.44	2.47	2.55
	CI	[7.15; 8.85]	[8.67; 10.39]	[8.15; 9.38]
Total	<i>n</i>	67	76	143
	<i>M</i>	7.06	8.76	7.97
	<i>SD</i>	2.60	2.37	2.62
	CI	[6.43; 7.69]	[8.22; 9.31]	[7.53; 8.40]

Note: *n* = sample size; *M* = mean; *SD* = standard deviation; CI = confidence interval; *k* = number of items

The ANOVA on video-dependent items showed no interaction between delivery mode and proficiency, as reflected in Table 4.13. The main effect of delivery mode was significant. The video-based mode was easier than the audio-only mode (see Table 4.12, last column), with η^2 of .11 (medium effect size). The significant main effect of proficiency shows that higher test-takers scored higher on the video-dependent items than lower test-takers, which was expected.

Table 4.13

Two-Way Factorial ANOVA on 14 Video-Dependent Items

Source	<i>df</i>	<i>F</i>	<i>p</i>	η^2
Delivery mode	1	17.35	< .001*	.11
Proficiency	1	20.49	< .001*	.13
Mode*Proficiency	1	0.44	.510	< .01
Error	139			

Note: *df* = degrees of freedom; *F* = F-statistic; *significant at α of .05; η^2 = partial eta squared; *N* = 143

The results of the second ANOVA did not support the researcher’s hypothesis that video-dependent items would be easier for higher-level participants but harder for lower-

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

level participants. These effects of mode were expected to cancel out, yielding no overall effect of mode on video-dependent items. This expectation was also not supported.

Instead, the video-dependent items were easier with videos than without.

ANOVA #3 on video-dependent local items. The video-based scores on video-dependent local items seemed to be consistently higher than the audio-only scores within both proficiency categories, suggesting no mode-delivery interaction (see Table 4.14). Overall, the video-based mode yielded higher scores than the audio-based mode, as reflected in the last column of Table 4.14.

Table 4.14

Descriptive Statistics for 7 Video-Dependent Local Items

Delivery mode		Proficiency		Total (<i>k</i> = 7)
		Lower	Higher	
Audio-only	<i>n</i>	33	42	75
	<i>M</i>	3.36	4.14	3.80
	<i>SD</i>	1.69	1.34	1.54
	CI	[2.76; 3.96]	[3.73; 4.56]	[3.45; 4.15]
Video-based	<i>n</i>	34	34	68
	<i>M</i>	4.09	4.85	4.47
	<i>SD</i>	1.56	1.33	1.49
	CI	[3.54; 4.63]	[4.39; 5.32]	[4.11; 4.83]
Total	<i>n</i>	67	76	143
	<i>M</i>	3.73	4.46	4.12
	<i>SD</i>	1.66	1.37	1.55
	CI	[3.33; 4.14]	[4.15; 4.77]	[3.86; 4.38]

Note: *n* = sample size; *M* = mean; *SD* = standard deviation; CI = confidence interval; *k* = number of items

The ANOVA on video-dependent local items showed no interaction between delivery mode and proficiency, as reflected in Table 4.15. The main effect of delivery mode was significant. The video-based mode was easier than the audio-only mode for local items that were video-dependent (see Table 4.14). The effect size was moderate ($\eta^2 = .06$). The significant main effect of proficiency shows that higher-proficiency test-takers outperformed lower-proficiency test-takers, as expected.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Table 4.15

Two-Way Factorial ANOVA on 7 Video-Dependent Local Items

Source	<i>df</i>	<i>F</i>	<i>p</i>	η^2
Delivery mode	1	8.34	.005*	.06
Proficiency	1	9.66	.002*	.07
Mode*Proficiency	1	0.00	.977	< .01
Error	139			

Note: *df* = degrees of freedom; *F* = F-statistic; * = significant at α of .05; η^2 = partial eta squared; *N* = 143

The third ANOVA did not support the hypothesis that the video-dependent local items would be easier for higher-level but harder for lower-level test-takers, as there was no interaction. These hypothesized contrasting effects were expected to cancel out, yielding no overall effect of mode. No support was found for this expectation. Again, video-dependent local items were easier for both proficiency groups.

ANOVA #4 on video-dependent global items. Video-dependent global items were easier in the video-based mode than in the audio-only mode irrespective of proficiency, as shown by higher video-based scores in the first two rows in Table 4.16.

Table 4.16

Descriptive Statistics for 7 Video-Dependent Global Items

Delivery mode		Proficiency		Total (<i>k</i> = 7)
		Lower	Higher	
Audio-only	<i>n</i>	33	42	75
	<i>M</i>	2.73	4.00	3.44
	<i>SD</i>	1.23	1.40	1.46
	CI	[2.29; 3.16]	[3.56; 4.44]	[3.10; 3.78]
Video-based	<i>n</i>	34	34	68
	<i>M</i>	3.91	4.68	4.29
	<i>SD</i>	1.26	1.59	1.48
	CI	[3.47; 4.35]	[4.12; 5.23]	[3.94; 4.65]
Total	<i>n</i>	67	76	143
	<i>M</i>	3.33	4.30	3.85
	<i>SD</i>	1.38	1.52	1.53
	CI	[2.99; 3.66]	[3.96; 4.65]	[3.59; 4.10]

Note: *n* = sample size; *M* = mean; *SD* = standard deviation; CI = confidence interval; *k* = number of items

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

The ANOVA on video-dependent global items showed no interaction between delivery mode and proficiency, as reflected in Table 4.17. The main effect of delivery mode was significant. The video-based mode was easier than the audio-only mode (see Table 4.16, last column). The effect size was moderate ($\eta^2 = .10$). The significant main effect of proficiency shows that higher test-takers expectedly outperformed lower test-takers on the video-dependent global items.

Table 4.17

Two-Way Factorial ANOVA on 7 Video-Dependent Global Items

Source	<i>df</i>	<i>F</i>	<i>p</i>	η^2
Delivery mode	1	16.10	< .001*	.10
Proficiency	1	19.30	< .001*	.12
Mode*Proficiency	1	1.20	.275	.01
Error	139			

Note: *df* = degrees of freedom; *F* = F-statistic; * = significant at α of .05; η^2 = partial eta squared; *N* = 143

The results of the fourth ANOVA did not support the researcher’s hypothesis that the video-dependent global items would be easier for higher-level test-takers but might be harder for lower-level test-takers, as there was no interaction. The hypothesis of no overall effect of mode was also rejected. Once again, video-dependent global items were easier for both groups.

ANOVA #5 on video-independent items. Next, video-independent items were examined in three analyses of video-dependent items overall (ANOVA #5), video-dependent global (ANOVA #6), and local items (ANOVA #7).

The video-based scores on video-independent items seemed to be similar to audio-only scores across both proficiency levels, suggesting no interaction between mode and proficiency (see Table 4.18). Overall, the video-based and audio-only modes yielded comparable total scores on video-independent items (Table 4.18, last column).

Table 4.18

Descriptive Statistics for 10 Video-Independent Items

Delivery mode		Proficiency		Total (<i>k</i> = 10)
		Lower	Higher	
Audio-only	<i>n</i>	33	42	75
	<i>M</i>	3.64	4.81	4.29
	<i>SD</i>	1.50	1.67	1.69
	CI	[3.11; 4.17]	[4.29; 5.33]	[3.90; 4.68]
Video-based	<i>n</i>	34	34	68
	<i>M</i>	3.44	5.24	4.34
	<i>SD</i>	2.03	2.35	2.36
	CI	[2.73; 4.15]	[4.42; 6.02]	[3.77; 4.91]
Total	<i>n</i>	67	76	143
	<i>M</i>	3.54	5.00	4.31
	<i>SD</i>	1.78	2.00	2.03
	CI	[3.10; 3.97]	[4.54; 5.46]	[3.98; 4.65]

Note: *n* = sample size; *M* = mean; *SD* = standard deviation; CI = confidence interval; *k* = number of items

The ANOVA on video-independent items showed no interaction between delivery mode and proficiency, as reflected in Table 4.19. The main effect of delivery mode was not significant. The video-based and audio-only modes were equally difficult, with a trivial effect size ($\eta^2 < .01$). The significant main effect of proficiency shows that higher test-takers outperformed lower test-takers on the video-independent items, as expected.

Table 4.19

Two-Way Factorial ANOVA on 10 Video-Independent Items

Source	<i>df</i>	<i>F</i>	<i>p</i>	η^2
Delivery mode	1	0.13	.719	< .01
Proficiency	1	21.48	< .001*	.13
Mode*Proficiency	1	0.00	.334	< .01
Error	139			

Note: *df* = degrees of freedom; *F* = F-statistic; *significant at α of .05; η^2 = partial eta squared; *N* = 143

The results of the fifth ANOVA supported the researcher’s hypothesis that video-independent items would have similar difficulty in the video-based mode and in the audio-based mode within each proficiency level.

ANOVA #6 on video-independent local items. Overall, the video-based and audio-only modes yielded comparable total scores on five video-independent local items, as shown by row totals in Table 4.20.

Table 4.20

Descriptive Statistics for 5 Video-Independent Local Items

Delivery mode		Proficiency		Total (<i>k</i> = 5)
		Lower	Higher	
Audio-only	<i>n</i>	33	42	75
	<i>M</i>	1.76	2.57	2.21
	<i>SD</i>	1.17	1.09	1.19
	CI	[1.34; 2.17]	[2.23; 2.91]	[1.94; 2.49]
Video-based	<i>n</i>	34	34	68
	<i>M</i>	1.50	2.85	2.32
	<i>SD</i>	1.25	1.44	1.44
	CI	[1.36; 2.23]	[2.53; 3.35]	[1.98; 2.67]
Total	<i>n</i>	67	76	143
	<i>M</i>	1.78	2.70	2.27
	<i>SD</i>	1.20	1.26	1.31
	CI	[1.48; 2.07]	[2.41; 2.98]	[2.05; 2.48]

Note: *n* = sample size; *M* = mean; *SD* = standard deviation; CI = confidence interval; *k* = number of items

The ANOVA on video-independent local items showed no interaction between delivery mode and proficiency, as reflected in Table 4.21. The main effect of delivery mode was not significant. The video-based and audio-only modes were equally difficult, with a trivial effect size ($\eta^2 < .01$). The significant effect of proficiency was expected.

Table 4.21

Two-Way Factorial ANOVA on 5 Video-Independent Local Items

Source	<i>df</i>	<i>F</i>	<i>p</i>	η^2
Delivery mode	1	0.59	.445	< .01
Proficiency	1	20.34	< .001*	.13
Mode*Proficiency	1	0.35	.334	< .01
Error	139			

Note: *df* = degrees of freedom; *F* = F-statistic; *significant at α of .05; η^2 = partial eta squared; *N* = 143

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

The results of the sixth ANOVA supported the researcher’s hypothesis that video-independent local items would have similar difficulty in the video-based and the audio-based modes within each proficiency category.

ANOVA #7 on video-independent global items. Video-independent global items were equally difficult with videos and without, as shown by row totals in Table 4.22. Video-independent global items were slightly harder for lower-level test-takers with videos. In contrast, they were somewhat easier with videos for higher-level test-takers. These differences seemed to cancel out in the total scores by mode, with the video-based and audio-only modes having similar total scores (see Table 4.22, last column).

Table 4.22

Descriptive Statistics for 5 Video-Independent Global Items

Delivery mode		Proficiency		Total (<i>k</i> = 5)
		Lower	Higher	
Audio-only	<i>n</i>	33	42	75
	<i>M</i>	1.88	2.24	2.08
	<i>SD</i>	1.08	1.12	1.11
	CI	[1.49; 2.26]	[1.89; 2.59]	[1.82; 2.34]
Video-based	<i>n</i>	34	34	68
	<i>M</i>	1.65	2.38	2.01
	<i>SD</i>	1.23	1.35	1.33
	CI	[1.22; 2.08]	[1.91; 2.85]	[1.69; 2.34]
Total	<i>n</i>	67	76	143
	<i>M</i>	1.76	2.30	2.05
	<i>SD</i>	1.16	1.22	1.22
	CI	[1.48; 2.04]	[2.02; 2.58]	[1.85; 2.25]

Note: *n* = sample size; *M* = mean; *SD* = standard deviation; CI = confidence interval; *k* = number of items

The ANOVA on video-independent global items showed no interaction between delivery mode and proficiency, as reflected in Table 4.23. The main effect of delivery mode was not significant. The video-based and audio-only modes were equally difficult,

with a trivial effect size ($\eta^2 < .01$). The significant main effect of proficiency shows that higher test-takers outperformed lower test-takers on the video-independent global items.

Table 4.23

Two-Way Factorial ANOVA on 5 Video-Independent Global Items

Source	<i>df</i>	<i>F</i>	<i>p</i>	η^2
Delivery mode	1	0.05	.828	< .01
Proficiency	1	7.41	.007*	.05
Mode*Proficiency	1	0.88	.351	.01
Error	139			

Note: *df* = degrees of freedom; *F* = F-statistic; *significant at α of .05; η^2 = partial eta squared; *N* = 143

The results of the seventh ANOVA supported the researcher’s hypothesis that video-independent global items would have similar difficulty in the video-based and the audio-based modes, regardless of proficiency.

Summary. The classical analysis results largely showed that content-rich videos had a considerable impact on listening comprehension difficulty. The video-based mode was easier for test-takers on the ALC test as a whole. In terms of items, video-dependent items were easier with videos than without. There was no such effect on video-independent items, as expected. Regarding the role of proficiency, the classical analysis did not show that lower- and higher-proficiency test-takers were differently affected by mode. The hypothesis of the video-based mode being easier for higher-level test-takers but harder for lower-level test-takers was not supported. Rather, the results indicated that videos helped both proficiency groups to the same extent. Finally, item type (i.e., local vs. global) did not play a role in the effect of delivery mode on listening comprehension.

Research question 1.1: Rasch Analysis. To have more precise estimates of the mode effect and the role of proficiency in it, Rasch analysis was run. It was based on the sample with the 24 repeaters (*N* = 167). A one-parameter Rasch model was set on the

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

following six facets: Test-taker (1-167), mode (audio-only and video-based), proficiency (lower and higher), item video-dependence (video-dependent and video-independent), item type (local and global), and individual items (1-24). Six analyses were run overall, starting with (1) the facet report for mode. Next, between-facets interactions were conducted to detect the following moderating effects on delivery mode: (2) proficiency, (3) video-dependence, (4) video-dependence and proficiency, (5) video-dependence and item type, and (6) video-dependence, proficiency, and item type. The structure of the Rasch Facets specification file is given in Appendix K.

Delivery mode. Table 4.24 displays the facet report for delivery mode. The video-based mode was slightly easier than the audio-only mode. The logit difficulty values (interpretation: the lower, the easier) for the audio-only and video-based groups were $M = 0.07$ ($SE=0.05$) and $M = -0.07$ ($SE=0.05$) respectively. As indicated by the chi-square statistics, $\chi^2(1) = 3.80$, separation index of 1.68, and $p = .05$, this difference approached statistical significance.

Table 4.24

Rasch Measurement Report for Delivery Mode

	Difficulty logit	Model's S.E.	Infit Mean Square	Infit Z
Audio-only ($n = 79$)	0.07	0.05	1.02	1.00
Video-based ($n = 88$)	-0.07	0.05	0.98	-1.10
<i>M</i>	0.00	0.05	1.00	0.00
<i>SD</i>	0.10	0.00	0.03	1.60

Note: n = sample size; M = mean; SD = standard deviation; S.E. = standard error.
Reliability = .74; Separation Index = 1.68; Fixed $\chi^2(1) = 3.80$, $p = .05$; $N = 167$.

The results of this analysis did not support the researcher's hypothesis of no difference in difficulty between the video-based and the audio-only modes (due to expected interaction with proficiency).

Proficiency and mode. Next, the proficiency-mode interaction analysis revealed that proficiency did not account for mode-based differences in difficulty. Table 4.25 shows the following information for each of the proficiency levels: listening difficulty when content-rich videos were absent (audio-only target measure) or present (video-based target measure), target contrast (difference between the target measures), joint standard errors, *t*-statistics, Welch degrees of freedom, and significance of the contrast. The video-based version was somewhat easier for higher-level test-takers but harder for lower-level test-takers. However, these differences did not reach statistical significance.

Table 4.25

Rasch Interaction: Proficiency in the Mode Effect

Proficiency	Target measure (S.E.)		Target contrast	Joint S.E.	<i>t</i>	Welch <i>df</i>	<i>p</i>
	Audio-only	Video-based					
Lower (<i>n</i> = 79)	-0.05 (0.07)	0.05 (0.07)	-0.11	0.10	-1.06	1893	.29
Higher (<i>n</i> = 88)	0.04 (0.06)	-0.05 (0.07)	0.10	0.10	0.98	2043	.33

Note: *n* = sample size; S.E. = Standard Error; *df* = degrees of freedom; *N* = 167.

The results of this Rasch bias/interaction analysis failed to support the researcher’s hypothesis that video-based ALC test would be easier for higher-level test-takers but harder for the lower-level test-takers relative to the audio-only mode.

Video-dependence and mode. Video-dependent items were easier in the video-based mode than in the audio-only mode, $t(2309) = 2.22, p = .03 < .05$ (see Table 4.26). In contrast, video-independent items were significantly harder in the video-based mode, $t(1651) = -2.63, p = .01 < .05$.

The results of this Rasch bias/interaction analysis did not support the researcher’s hypothesis that video-dependent items would not differ in difficulty across modes, when

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

analyzed without regard to proficiency. The hypothesis of unaffected video-independent items was also rejected.

Table 4.26

Rasch Interaction: Video-Dependence in the Mode Effect

Items	Target measure (S.E.)		Target contrast	Joint S.E.	<i>t</i>	Welch <i>df</i>	<i>p</i>
	Audio-only	Video-based					
Video-dependent (<i>k</i> = 14)	-0.10 (0.06)	-0.30 (0.07)	0.21	0.09	2.22	2309	.03*
Video-independent (<i>k</i> = 10)	0.06 (0.07)	0.34 (0.08)	-0.28	0.11	-2.63	1651	.01*

Note: *k* = number of test items; S.E. = Standard Error; * = significance at $\alpha = .05$; *df* = degrees of freedom; *N* = 167

Video-dependence, proficiency, and mode. Video-dependent items’ difficulty for both lower and higher proficiency levels was not significantly different for the video-based versus the audio-only mode (see Table 4.27). Video-independent items were significantly harder for lower-level test-takers in the video-based mode, $t(787) = -3.25, p = .001$, but not for higher-level participants.

Table 4.27

Rasch Interaction: Video-Dependence and Proficiency for Mode

Items	Proficiency	Target measure (S.E.)		Target contrast	Joint S.E.	<i>t</i>	<i>df</i>	<i>p</i>
		Audio-only	Video-based					
Video-dependent	Lower (<i>n</i> = 79)	-0.12 (0.09)	-0.30 (0.09)	0.17	0.13	1.31	1103	.190
	Higher (<i>n</i> = 88)	-0.07 (0.09)	-0.31 (0.10)	0.24	0.13	1.79	1186	.073
Video-independent	Lower (<i>n</i> = 79)	-0.05 (0.11)	0.47 (0.11)	-0.52	0.16	-3.25	787	.001*
	Higher (<i>n</i> = 88)	0.13 (0.10)	0.21 (0.11)	-0.08	0.15	-0.55	852	.586

Note: S.E. = Standard Error; *significance at $\alpha = .05$; *df* = Welch degrees of freedom; *N* = 167

The results of this Rasch bias/interaction analysis did not support the researcher’s hypothesis that video-dependent items would be easier in the video-based mode for higher-level test-takers but harder for the lower-level test-takers. The hypotheses of video-independent items being unaffected by mode and proficiency was also rejected.

Video-dependence, item type, and mode. Video-dependent items of either local or global type did not differ in difficulty across the audio-only and video-based modes (see Table 4.28). There was no difference in the difficulty of video-independent local items across the delivery modes. In contrast, video-independent global items were significantly harder with videos than with audio-only, $t(824) = -2.35, p = .02$.

The results of this Rasch bias/interaction analysis supported the researcher’s hypotheses that both local and global video-dependent items would not differ in difficulty across the modes. The hypotheses of both local and global video-independent items being unaffected by mode was rejected.

Table 4.28

Rasch Interaction: Video-Dependence and Item Type for Mode

Items	Item type	Target measure (S.E.)		Target contrast	Joint S.E.	<i>t</i>	Welch <i>df</i>	<i>p</i>
		Audio-only	Video-based					
Video-dependent	Local	-0.16 (0.09)	-0.34 (0.10)	0.18	0.13	1.36	1153	.17
	Global	-0.03 (0.09)	-0.26 (0.10)	0.23	0.13	1.78	1154	.08
Video-independent	Local	0.04 (0.10)	0.24 (0.11)	-0.21	0.15	-1.37	825	.17
	Global	0.07 (0.11)	0.43 (0.11)	-0.36	0.15	-2.35	824	.02*

Note: S.E. = Standard Error; * = significance at $\alpha = .05$; *df* = degrees of freedom; *N* = 167

Video-dependence, proficiency, item type, and mode. Finally, the last interaction was run to see if proficiency and item type accounted for mode-based differences in difficulty for video-dependent and video-independent items found previously (see Table 4.26). For the video-dependent subset, lower- and higher-level test-takers did not perform differently on either local or global items, across the audio-only and video-based modes (see Table 4.29). For the video-independent subset, lower-level test-takers performed worse on global items under the video-based condition, $t(392) = -2.68, p = .01$. Higher-

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

level test-takers did not perform differently on either local or global video-independent items by mode.

Table 4.29

Rasch Interaction: Video-Dependence, Proficiency, Item Type for Mode

Items, Proficiency	Item type	Target measure (S.E.)		Target contrast	Joint S.E.	<i>t</i>	Welch <i>df</i>	<i>p</i>
		Audio-only	Video-based					
VDEP, Lower	Local	-0.26 (0.13)	-0.30 (0.13)	0.04	0.19	0.22	550	.83
	Global	-0.01 (0.13)	-0.29 (0.13)	0.31	0.19	1.64	550	.10
VDEP, Higher	Local	-0.08 (0.12)	-0.40 (0.14)	0.31	0.19	1.67	590	.10
	Global	-0.06 (0.12)	-0.22 (0.14)	0.16	0.18	0.87	593	.38
VIND, Lower	Local	0.03 (0.16)	0.46 (0.16)	-0.43	0.22	-1.92	392	.06
	Global	-0.12 (0.16)	0.48 (0.16)	-0.60	0.23	-2.68	392	.01*
VIND, Higher	Local	0.04 (0.14)	0.03 (0.15)	0.01	0.21	0.04	424	.97
	Global	0.22 (0.14)	0.39 (0.15)	-0.17	0.21	-0.81	425	.42

Note: VDEP = video-dependent items; VIND = video-independent items; S.E. = Standard Error; *df* = degrees of freedom; * significance at $\alpha = .05$; $N = 167$

The results of the final Rasch bias/interaction analysis did not support the researcher’s hypothesis that, for lower-level test-takers, video-dependent local and global items would be harder with videos than with audio, but, for higher-level test-takers, they would be easier in the video-based mode. The hypothesis about no effect of mode on both global and local video-independent items was supported only at the higher proficiency.

Summary. The Rasch analysis results showed that content-rich videos had an effect on listening comprehension difficulty. The video-based mode was found to be easier for test-takers on the video-dependent items and on the ALC test overall. This was similar to the classical analysis results described in the previous section.

Rasch analysis had the capacity to further uncover effects of mode, proficiency, video-dependence, and item type. Table 4.30 compares the results for research question 2.1 from the classical analysis and the Rasch analysis for each dependent and independent variable. Unlike the classical analyses, the Rasch analysis showed that

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

video-independent items were harder in the video-based mode than in the audio-only mode. This effect was moderated by item type and proficiency. Specifically, video-independent global items were significantly harder for lower-level test-takers with videos than with audio-only.

Table 4.30

Comparison of Classical and Rasch Analyses Results for RQ 1.1

DV	IV	Classical	Rasch
all items	Mode	Easier	Easier
	Proficiency	-	-
video-dependent	Mode	Easier	Easier
	Proficiency	-	-
video-dependent local	Mode	Easier	-*
	Proficiency	-	-
video-dependent global	Mode	Easier	-*
	Proficiency	-	-
video-independent items	Mode	-	Harder*
	Proficiency	-	Harder for lower proficiency*
video-independent local	Mode	-	-
	Proficiency	-	-
video-independent global	Mode	-	Harder*
	Proficiency	-	Harder for lower proficiency*

Note: DV = dependent variable; IV = independent variable; - no effect; * a Rasch analysis result that is different from the parallel classical analysis result

Research question 1.2. Research question 1.2 asked: Is academic listening comprehension difficulty at the item level affected by delivery mode, listening proficiency, item video-dependence, and/or item type? The effects of delivery mode and proficiency on each individual item’s difficulty could be statistically detected by running two Rasch bias/interaction analyses, one for the effect of mode and the other for the joint effect of mode and proficiency. Effects of video-dependence and item type were examined descriptively by looking for patterns in the outputs of these interactions.

Delivery mode. The first interaction analysis revealed significant effects of mode on three ALC items. The 24 interactions are displayed in Table 4.31 below, with the indication of video-dependence, item number, type, difficulty logits by mode, the target contrast between them (a negative contrast means that the item is harder in the video-based mode), joint standard errors, *t*-statistics, Welch degrees of freedom, and *p*-values. According to the table, video-dependent local item 19 was harder in the video-based mode but video-dependent local item 22 was easier in the video-based mode. Video-independent global item 24 was harder in the video-based mode than in the audio-only mode. All the three items were part of the Exoplanets testlet.

Descriptively, Table 4.31 reveals that 10 out of the 14 video-dependent items were easier in the video-based mode than in the audio-only mode. Half of them were local (2, 8, 9, 15, 22) and the other half global (4, 7, 13, 20, 23). The remaining four video-dependent items were harder in the video-based mode than in the audio-only mode. Half of them were local (5, 19), and the other half global (12, 18).

Seven of the 10 video-independent items were harder in the video-based mode than in the audio-only mode. About half of them were local (14, 16, 21), and the others global (3, 6, 10, 24). The remaining three items were easier in the video-based mode than in the audio-only mode. Two of them were local (1, 11) and the other one global (17).

Another potential pattern was revealed for global main idea items (6, 12, 18, 24). They all were harder in the video-based mode than in the audio-only mode, regardless of video-dependence. For item 24, this pattern reached statistical significance.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Table 4.31

ALC Test Items: Rasch Item-Mode Interactions

Video-dependence	Item	Type	Target measure		Target contrast	Joint S.E.	<i>t</i>	Welch <i>d.f.</i>	<i>p</i>
			Audio-only	Video-based					
VDEP (<i>k</i> = 14)	Item #2	LOC	-0.37	-0.70	0.32	0.36	0.91	162	.363
	Item #4	GLO	0.42	0.08	0.34	0.33	1.02	163	.309
	Item #5	LOC	0.00	0.42	-0.42	0.33	-1.25	163	.212
	Item #7	GLO	0.00	-0.10	0.10	0.33	0.29	163	.773
	Item #8	LOC	-0.90	-1.17	0.27	0.39	0.69	161	.489
	Item #9	LOC	1.22	0.77	0.45	0.35	1.28	163	.201
	Item #12	GLO	0.10	0.31	-0.21	0.33	-0.62	163	.533
	Item #13	GLO	0.26	-0.34	0.60	0.34	1.78	162	.078
	Item #15	LOC	-0.15	-0.42	0.27	0.34	0.78	162	.438
	Item #18	GLO	-1.77	-1.57	-0.19	0.44	-0.43	163	.667
	Item #19	LOC	-0.48	0.25	-0.73	0.34	-2.16	163	.033*
	Item #20	GLO	0.36	-0.04	0.41	0.33	1.21	163	.228
	Item #22	LOC	0.32	-0.92	1.24	0.36	3.42	160	.001*
	Item #23	GLO	0.42	0.02	0.40	0.33	1.19	163	.235
VIND (<i>k</i> = 10)	Item #1	LOC	0.53	0.45	0.09	0.34	0.26	163	.796
	Item #3	GLO	-0.28	0.22	-0.50	0.33	-1.49	163	.137
	Item #6	GLO	0.66	0.96	-0.30	0.36	-0.84	163	.404
	Item #10	GLO	0.54	0.76	-0.22	0.35	-0.64	163	.524
	Item #11	LOC	0.09	-0.31	0.41	0.33	1.21	163	.228
	Item #14	LOC	-0.12	0.45	-0.56	0.33	-1.68	163	.094
	Item #16	LOC	-0.48	0.10	-0.58	0.33	-1.74	163	.083
	Item #17	GLO	-0.28	-0.48	0.20	0.33	0.60	163	.548
Item #21	LOC	0.26	0.63	-0.37	0.34	-1.09	163	.277	
Item #24	GLO	-0.39	0.57	-0.96	0.34	-2.84	163	.005*	

Note: VDEP = video-dependent, VIND = video-independent, LOC = local, GLO = global, S.E. = Standard Error, *significant at .05 level; *N* = 167

Mode and proficiency. The second Rasch bias/interaction analysis revealed significant effects of mode in both lower and higher proficiency groups.

Lower proficiency. Within the lower-proficiency category, four items displayed significant effects of mode (10, 14, 19, 24). They all were video-independent except for item #19. All the items were harder in the video-based mode than in the audio-only mode, as reflected in Table 4.32.

Statistical significance aside, Table 4.32 shows that 10 out of 14 video-dependent items were easier with videos than with audio-only, with no apparent role of item type.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Five of these items were local (2, 5, 8, 15, 22) and the other five global (4, 7, 13, 20, 23).

The remaining four items were harder in the video-based mode. Again, they were half local (9, 19) and half global (12, 18).

Table 4.32

Effects of Mode on Individual Items' Difficulties at Lower Proficiency

Video-dependence	Item	Type	Target measure		Target contrast	Joint S.E.	<i>t</i>	Welch <i>d.f.</i>	<i>p</i>
			Audio-only	Video-based					
VDEP (<i>k</i> = 14)	Item #2	LOC	-0.66	-1.18	0.51	0.52	0.99	76	.324
	Item #4	GLO	0.64	0.18	0.47	0.50	0.94	76	.352
	Item #5	LOC	0.25	0.06	0.19	0.48	0.39	75	.700
	Item #7	GLO	-0.21	-0.27	0.06	0.48	0.14	76	.893
	Item #8	LOC	-0.66	-0.76	0.10	0.50	0.19	76	.848
	Item #9	LOC	0.76	0.87	-0.12	0.51	-0.23	76	.821
	Item #12	GLO	0.02	0.29	-0.27	0.48	-0.56	76	.577
	Item #13	GLO	0.14	-0.62	0.76	0.49	1.56	76	.124
	Item #15	LOC	0.02	-0.16	0.18	0.48	0.37	76	.710
	Item #18	GLO	-1.85	-1.62	-0.23	0.59	-0.38	76	.715
	Item #19	LOC	-0.66	0.52	-1.18	0.48	-2.45	76	.017*
	Item #20	GLO	0.64	0.06	0.58	0.50	1.16	76	.249
	Item #22	LOC	-0.10	-0.76	0.66	0.49	1.35	76	.182
	Item #23	GLO	0.78	0.06	0.72	0.51	1.42	76	.159
VIND (<i>k</i> = 10)	Item #1	LOC	1.00	1.35	-0.34	0.60	-0.58	76	.566
	Item #3	GLO	-0.48	0.14	-0.61	0.48	-1.28	76	.206
	Item #6	GLO	0.26	0.63	-0.37	0.52	-0.72	76	.476
	Item #10	GLO	0.26	1.41	-1.15	0.57	-2.02	76	.047*
	Item #11	LOC	0.24	-0.32	0.56	0.49	1.13	76	.261
	Item #14	LOC	-0.25	0.88	-1.13	0.51	-2.24	76	.028*
	Item #16	LOC	-0.70	0.13	-0.84	0.48	-1.75	76	.084
	Item #17	GLO	-0.48	-0.43	-0.05	0.48	-0.10	76	.922
	Item #21	LOC	0.11	0.49	-0.38	0.50	-0.76	76	.452
	Item #24	GLO	-0.36	0.63	-1.00	0.50	-1.99	76	.050*

Note: VDEP = video-dependent, VIND = video-independent, LOC = local, GLO = global, S.E. = Standard Error, *significant at .05 level; *n* = 79

Out of the 10 video-independent items, nine were harder with the videos than with audio-only. Item type seemed to have no role in this relationship. Four of these items were local (1, 14, 16, 21) and the other five global (3, 6, 10, 17, 24). One video-independent item was easier in the video-based mode (item 11, local).

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Higher proficiency. For higher-level test-takers, the Rasch bias/interaction analysis revealed significant contrasts for three items (see Table 4.33). Two of these items were video-dependent, one of which was harder (item 5, local) and the other easier (item 22, local) with videos than with audio-only. The remaining item was video-independent and was harder in the video-based mode (item 24, global). The role of item type in these mode-item interaction effects was inconclusive.

Table 4.33

Effects of Mode on Individual Items' Difficulties at Higher Proficiency

Video-dependence	Item	Type	Target measure		Target contrast	Joint S.E.	<i>t</i>	Welch <i>d.f.</i>	<i>p</i>
			Audio-only	Video-based					
VDEP (<i>k</i> = 14)	Item #2	LOC	-0.11	-0.18	0.07	0.48	0.15	82	.880
	Item #4	GLO	0.26	-0.03	0.29	0.47	0.63	82	.533
	Item #5	LOC	-0.21	0.79	-0.99	0.47	-2.13	83	.036*
	Item #7	GLO	0.17	0.09	0.08	0.46	0.16	83	.871
	Item #8	LOC	-1.19	-2.00	0.80	0.73	1.09	78	.277
	Item #9	LOC	1.54	0.67	0.87	0.48	1.83	84	.071
	Item #12	GLO	0.17	0.33	-0.16	0.46	-0.35	83	.724
	Item #13	GLO	0.35	-0.03	0.38	0.47	0.82	82	.412
	Item #15	LOC	-0.31	-0.78	0.47	0.53	0.89	81	.375
	Item #18	GLO	-1.66	-1.51	-0.16	0.67	-0.23	83	.817
	Item #19	LOC	-0.31	-0.05	-0.26	0.48	-0.54	83	.588
	Item #20	GLO	0.17	-0.16	0.33	0.47	0.69	82	.492
	Item #22	LOC	0.65	-1.16	1.81	0.55	3.28	78	.002*
	Item #23	GLO	0.17	-0.03	0.20	0.47	0.43	82	.672
VIND (<i>k</i> = 10)	Item #1	LOC	0.27	-0.30	0.57	0.47	1.22	82	.226
	Item #3	GLO	-0.12	0.29	-0.42	0.46	-0.91	83	.364
	Item #6	GLO	0.94	1.26	-0.32	0.49	-0.66	83	.513
	Item #10	GLO	0.73	0.29	0.44	0.47	0.94	83	.350
	Item #11	LOC	-0.01	-0.30	0.29	0.47	0.63	82	.533
	Item #14	LOC	-0.01	0.06	-0.07	0.46	-0.15	83	.880
	Item #16	LOC	-0.29	0.06	-0.35	0.46	-0.76	83	.448
	Item #17	GLO	-0.12	-0.54	0.42	0.47	0.89	82	.377
Item #21	LOC	0.36	0.75	-0.39	0.46	-0.86	83	.394	
Item #24	GLO	-0.41	0.52	-0.93	0.46	-2.02	83	.047*	

Note: VDEP = video-dependent, VIND = video-independent, LOC = local, GLO = global, S.E. = Standard Error, *significant at .05 level; *n* = 88

Examining Table 4.33 descriptively, we find that 10 out of 14 video-dependent items were easier in the video-based mode than in the audio-only mode. There was no

role of item type in this trend. Half of these 10 items were local (2, 8, 9, 15, 22) and the other half global (4, 7, 13, 20, 23). In contrast, the remaining four video-dependent items were harder with the videos, including two local (5, 19) and two global items (12, 18).

Out of 10 video-independent items, six were harder with videos than with audio-only, regardless of item type. Half of these items were local (14, 16, 21) and the other half was global (3, 6, 24). The remaining four video-independent items were easier in the video-based mode, including two local (1, 11) and two global items (10, 17).

Comparison across proficiency levels. To finalize the descriptive analyses, the identified trends were compared across lower and higher proficiencies. Table 4.34 compares items that were harder or easier in the video-based mode by video-dependence and proficiency.

Regarding video-dependent items that were easier in the video-based mode, there is a noticeable overlap between the lower- and higher proficiency groups (see Table 4.34). This shows that items on which lower and higher test-takers performed better in the video-based mode were largely the same (2, 4, 7, 8, 13, 15, 20, 22, 23). Regarding the video-dependent items that were harder in the video-based mode, we see a similar trend (see Table 4.34). This shows that items on which lower and higher test-takers performed worse in the video-based mode were mostly the same (12, 18, 19).

For video-independent items that were easier in the video-based mode, there is little overlap between the lower and higher proficiency groups (i.e., item 11). This shows that lower-level test-takers more frequently performed better on video-independent items in the video-based mode than higher-level test-takers did. For video-independent items that were harder in the video-based mode, there is some overlap between the lower and

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

higher proficiency groups. This shows that lower and higher test-takers performed worse on the same seven items in the video-based mode (3, 6, 14, 16, 21, 24).

Table 4.34

Comparison of ALC Items' Difficulty in the Video-Based Mode

Video-dependence	Item #	Easier in the video-based mode		Harder in the video-based mode	
		Lower proficiency	Higher proficiency	Lower proficiency	Higher proficiency
VDEP	Item #2	X	X		
	Item #4	X	X		
	Item #5	X			X
	Item #7	X	X		
	Item #8	X	X		
	Item #9		X	X	
	Item #12			X	X
	Item #13	X	X		
	Item #15	X	X		
	Item #18			X	X
	Item #19			X	X
	Item #20	X	X		
	Item #22	X	X		
Item #23	X	X			
VIND	Item #1		X	X	
	Item #3			X	X
	Item #6			X	X
	Item #10		X	X	
	Item #11	X	X		
	Item #14			X	X
	Item #16			X	X
	Item #17		X	X	
	Item #21			X	X
	Item #24			X	X

Note: VDEP = video-dependent items; VIND = video-independent items; N = 167

Summary. It was hypothesized that each video-dependent item would be easier in the video-based mode than in the audio-only mode for higher-level students. The Rasch-based analyses for research question 2.2 did not support the researcher’s hypotheses. For lower-level students, an opposite effect was expected. While the descriptive analyses showed that most of the video-dependent items were easier with the videos, no role of proficiency and item type was found. As far as inferential statistics are concerned, the hypotheses also found little support. While some video-dependent items were found

significantly easier with the videos, others were harder. The numbers of these items was small, and no role of proficiency was discernable.

It was also hypothesized that each video-independent item would not be affected by delivery mode within either proficiency level and for either item type. While upheld for some items, this hypothesis was rejected for most items. The Rasch analysis revealed that some video-independent items were harder with the videos than without the videos. The descriptive analysis showed that most of the video-independent items were harder with the videos for both lower- and higher-level test-takers regardless of item type.

Research question 1.3. Research question 1.3 asked: Is academic listening difficulty related to viewing behavior and listening proficiency? The viewing behavior measure consisted of four identical questions asking test-takers to report their viewing behavior perceptions after each of the four ALC testlets as a score from 1 (did not watch the video) to 5 (watched all of the video). These four scores were combined into a composite score ranging from 1 to 20.

Using composite scores was justified by the following reasons. First, the four questions were designed to measure the same construct. Second, the scores on each question were positively correlated. Table 4.35 shows Spearman's correlations for each pair of testlets. The positive correlations indicated that the four testlet-based questions worked in the same direction, and, thus, would not be washed away in the composite score. Third, Cronbach's alpha internal-consistency reliability for the four questions was as high as 0.92, with the four item-total correlation indices being greater than 0.80. This shows that the four questions could be reliably combined in one score.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Table 4.35

Spearman’s Correlations for Viewing Behavior Ratings by Testlet

	VB on Homeostasis	VB on Food Tax	VB on Compassion	VB on Exoplanets
VB on Homeostasis	1.00	.68*	.69*	.65*
VB on Food Tax		1.00	.70*	.77*
VB on Compassion			1.00	.88*
VB on Exoplanets				1.00

VB = viewing behavior scores; * = significance at $\alpha = .01$

Table 4.36 displays sample sizes, means, standard deviations, and confidence intervals for viewing behavior on each testlet and overall. Higher composite scores indicated higher degree of viewing behavior. We see that lower-level test-takers watched videos slightly less attentively than higher-level test-takers across the testlets and overall. This is particularly noticeable for the Compassion and Exoplanets testlets. The viewing behavior mean scores by testlet ranged from 4.03 (Exoplanets) to 4.47 (Homeostasis), showing that test-takers watched each video most of the time.

Table 4.36

Descriptive Statistics for Viewing Behavior by Proficiency and Testlet

Proficiency		Testlet				Total
		Homeostasis	Food Tax	Compassion	Exoplanets	
Lower	<i>k</i>	1	1	1	1	4
	<i>n</i>	34	34	34	34	34
	<i>M</i>	4.26	4.06	3.76	3.74	15.82
	<i>SD</i>	0.99	1.18	1.35	1.36	4.39
	CI	[3.92; 4.61]	[3.65; 4.47]	[3.29; 4.24]	[3.26; 4.21]	[14.29; 17.36]
Higher	<i>k</i>	1	1	1	1	4
	<i>n</i>	34	34	34	34	34
	<i>M</i>	4.68	4.65	4.41	4.32	18.06
	<i>SD</i>	0.81	0.81	0.96	1.20	3.37
	CI	[4.40; 4.96]	[4.36; 4.93]	[4.08; 4.75]	[3.91; 4.74]	[16.89; 19.23]
Overall	<i>k</i>	1	1	1	1	4
	<i>n</i>	68	68	68	68	68
	<i>M</i>	4.47	4.35	4.09	4.03	16.94
	<i>SD</i>	0.92	1.05	1.21	1.30	4.04
	CI	[4.25; 4.69]	[4.10; 4.61]	[3.80; 4.38]	[3.71; 4.35]	[15.96; 17.92]

Note: *k* = number of items; *n* = sample size; *M* = mean; *SD* = standard deviation; CI = confidence interval

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

To answer research question 1.3, multiple regression was set to be used initially. However, the viewing behavior and ALC test scores were not linearly related, violating the fundamental assumption for linear regression. Instead, three Spearman’s Rank-Order correlation analyses were used, one for the lower-proficiency group, one for the higher-proficiency group, and one overall. The only assumption for Spearman’s correlation is having two ordinal, interval, or ratio variables. Viewing behavior composite ratings and ALC scores were assumed to be interval.

The results of the correlation analyses are given in Table 4.37. As shown in the table, correlation between ALC total scores and viewing behavior scores was not significant within either proficiency category. Overall, the correlation was weak but significant, $r = 0.29$, $p = 0.018 < 0.5$.

Table 4.37

Correlation Analyses for Viewing Behavior and ALC Test Scores

	ALC total score		
	Lower proficiency	Higher proficiency	Overall
<i>n</i>	34	34	68
Composite viewing behavior score	.22 ($p = .210$)	.06 ($p = .741$)	.29* ($p = .018$)

The results of the correlation analyses did not support the hypothesis that the lower-level test-takers’ viewing behavior would be negatively related to their ALC test scores while higher-level test-takers’ viewing behavior would positively relate to their ALC test scores. It was believed that these contrastive relationships would neutralize the overall correlation between viewing behavior and ALC test scores, regardless of proficiency. This hypothesis was also rejected. Instead, a moderate positive overall correlation was found.

Research Question 2

The second research question asked: Do stakeholders' perceptions lend support for using content-rich videos in the L2 academic listening assessment construct? This research question was subdivided into two subquestions. Research question 2.1 investigated test-takers' perceptions while research question 2.2 examined teachers' perceptions. This section provides statistical results for each of these two questions.

Research question 2.1. Research question 2.1 asked: Do delivery mode and listening proficiency affect test-takers' perceptions about listening difficulty, motivation, authenticity, and use of content-rich videos in tests? It was hypothesized that higher-level test-takers would perceive the video-based mode to be easier, more motivating, and more authentic. Also, test-takers in the video group were expected to favor the use of videos in L2 academic listening tests to a greater extent than test-takers in the audio-only group. Lower-level students were expected to have similar perceptions except for difficulty. It was hypothesized that lower-level test-takers would find the video-based mode more difficult than the audio-only mode.

Operationalization of variables. Difficulty, motivation, and authenticity were measured separately for each of the four ALC testlets. Taking difficulty as an example, one difficulty question was asked after test-takers listened to (or listened to and watched) the first testlet (i.e., Homeostasis.) The next difficulty question was identical to the first question, but was administered after the following Food Tax testlet. The next two difficulty questions came after Compassion and Exoplanets respectively. From these four testlet-based scores, a composite difficulty score was derived to estimate the overall test-

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

takers’ difficulty perceptions about the test. Composite scores for motivation and authenticity were derived analogously from the respective testlet-based ratings.

Using composite scores for difficulty, motivation, and authenticity was justified. The respective testlet-based component questions were theorized to measure the same constructs. For example, the four difficulty questions measured the construct “test-takers’ difficulty perceptions of the ALC test.” While each of the four scores estimated difficulty perceptions on a different component (testlet) of the ALC test, their summation provided an estimation of the overall test difficulty perceptions. Furthermore, the component scores were positively correlated, supporting the said justification. Table 4.38 shows the inter-correlation matrices for difficulty, motivation, and authenticity by ALC testlet. Although the correlation indices were somewhat low, they all were positive, showing that the component measures within each construct worked in the same direction across the four ALC testlets.

Table 4.38

Correlations for Difficulty, Motivation, and Authenticity Perceptions

Construct	Testlet	Homeostasis	Food Tax	Compassion	Exoplanets
Difficulty	Homeostasis	1.00	.53	.39	.43
	Food Tax		1.00	.57	.60
	Compassion			1.00	.55
	Exoplanets				1.00
Motivation	Homeostasis	1.00	.35	.32	.34
	Food Tax		1.00	.24	.18
	Compassion			1.00	.31
	Exoplanets				1.00
Authenticity	Homeostasis	1.00	.48	.37	.40
	Food Tax		1.00	.33	.33
	Compassion			1.00	.63
	Exoplanets				1.00

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Internal consistency reliability can also support the use of composite scores. Table 4.39 shows Cronbach's alpha reliability indices for difficulty, motivation, authenticity, and video use questions, along with the number of items, composite score ranges, and corrected item-total correlation ranges. We see that the overall reliability indices were adequate, considering the recommended cut-off of .70. This shows that items consistently measured the intended construct and could be combined for a composite score. The item-total correlations ($> .25$) show that items fit well into the measured constructs.

Table 4.39

Reliabilities for Difficulty, Motivation, and Authenticity Perceptions

Construct	k	Score range	Overall α	α by mode		Item-total correlation range
				Audio-only	Video-based	
Difficulty	4	1-24	.82	.80	.82	.51-.78
Motivation	4	1-24	.63	.69	.53	.24-.55
Authenticity	4	1-24	.71	.71	.70	.32-.63
Video use	3	1-18	.66	.66	.63	.31-.63

Each testlet-based question within each of the constructs of difficulty, motivation, and authenticity was measured on a 6-point ordinal scale (i.e. 1 – very easy, 6 – very difficult; 1 – very boring, 6 – very interesting; 1 – not realistic, 6 – very realistic). Therefore, the three corresponding composite scores ranged from 1 to 24, making it possible to treat each composite score as continuous (Schumacker & Lomax, 2004).

The construct use of videos in tests was measured differently from difficulty, motivation, and authenticity. There were no testlet-based questions. All the three questions (compare to four questions in the other constructs) appeared after the four ALC testlets. The video use composite score ranged from 1-18 and was also considered to be on a continuous scale (see Table 4.39).

Assumption check. The originally-intended multiple regression analysis was discarded due to violations of the linearity and multicollinearity assumptions. Instead, research question 2.1 was answered by running four separate ANOVAs. The ANOVAs examined the effects of mode and proficiency on test-takers' opinions about (1) difficulty, (2) motivation, (3) authenticity, and (4) the use of content-rich videos in listening tests. For each ANOVA, the following assumptions were checked: (a) independence of observations, (b) no significant outliers, (c) normality of the dependent variable's distribution for each combination of the groups of the independent variables, and (d) homogeneity of variance for each combination of the groups of the independent variables. Independence of observations was warranted by the design of the study. No significant outliers ($z \geq 3.29$) were found.

Normality was checked using Q-Q plots, skewness/kurtosis values, and Shapiro-Wilk's tests. Out of the 16 examined combinations of variables (i.e., $4 \times 2 \times 2$), three had a significant Shapiro-Wilk's statistic. Because the majority of the data was normally distributed and ANOVA is relatively robust against violations of normality, the normality assumption was assumed to be met for each of the perception constructs of difficulty, motivation, authenticity, and video use. The homogeneity-of-variance assumption was also met, as suggested by non-significant Levene's test statistics.

Difficulty perceptions. The video-based mode was perceived as slightly easier than the audio-only mode within each proficiency category and collectively (see Table 4.40), $M = 16.16$ out of 24, $SD = 4.22$ (audio-only) and $M = 14.59$ out of 24, $SD = 4.27$ (video-based). Recall that the higher ratings indicate higher difficulty. Being greater than the mid-point of 12, the means show that both groups perceived ALC lectures to be hard

in general, though to slightly different degrees. As seen in the last row of Table 4.40, higher- and lower-level learners were not different in their overall difficulty perceptions.

Table 4.40

Descriptive Statistics for Test-Takers' Difficulty Perceptions

Delivery mode		Proficiency		Total
		Lower	Higher	
Audio-only	<i>n</i>	33	42	75
	<i>M</i> (out of 24)	16.52	15.88	16.16
	<i>SD</i>	4.41	4.10	4.22
	CI	[14.95; 18.08]	[14.60; 17.16]	[15.19; 17.13]
Video-based	<i>n</i>	34	34	68
	<i>M</i> (out of 24)	15.06	14.12	14.59
	<i>SD</i>	4.10	4.43	4.27
	CI	[13.63; 16.49]	[12.57; 15.66]	[13.56; 15.62]
Total	<i>n</i>	67	76	143
	<i>M</i> (out of 24)	15.78	15.09	15.41
	<i>SD</i>	4.29	4.32	4.30
	CI	[14.73; 16.82]	[14.11; 16.08]	[14.70; 16.12]

The interaction term was not significant, as shown by the first ANOVA (see Table 4.41). The video-based mode was perceived to be easier compared to the audio-only mode (see the means in Table 4.40). However, the size of this effect was small ($\eta^2 = .04$). The main effect of proficiency was not significant, with a trivial effect size ($\eta^2 = .01$).

Table 4.41

Two-Way Factorial ANOVA on Test-Takers' Difficulty Perceptions

Source	<i>df</i>	<i>F</i>	<i>p</i>	η^2
Delivery mode	1	5.07	.026*	.04
Proficiency	1	1.21	.272	.01
Mode*Proficiency	1	0.05	.830	< .01
Error	139			

Note: *df* – degrees of freedom; *F* – F-statistic; η^2 – partial eta squared; *significant at $\alpha = .05$; *N* = 143

The first ANOVA results did not support the researcher's hypothesis that higher-level test-takers would find the video-based mode to be easier than the audio-only mode but lower-level test-takers would consider the video-based harder. These two effects were

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

expected to neutralize, with no overall effect of mode. This hypothesis was also rejected. However, the results were in accord with the learners’ performance on the ALC test.

Motivation perceptions. The two modes generated motivation ratings of similar magnitude, $M = 16.33$, $SD = 4.32$ (audio-only) and $M = 16.71$, $SD = 3.46$ (video-based), as shown in Table 4.42. This indicates that test-takers in both groups were equally interested in the listening lectures. Within each delivery mode, lower-level test-takers found passages less interesting than higher-level test-takers (see Table 4.42; higher score indicate more interest). Total scores for the lower and higher proficiency categories also reflected this trend ($M = 15.69$ out of 24, $SD = 4.16$ and $M = 17.24$, $SD = 3.58$ respectively). Being greater than the midpoint of $24/2 = 12$, these means showed that both proficiency groups considered ALC test lectures interesting on average, though to slightly different degrees.

Table 4.42

Descriptive Statistics for Test-Takers’ Motivation Perceptions

Delivery mode		Proficiency		Total
		Lower	Higher	
Audio-only	<i>n</i>	33	42	75
	<i>M</i> (out of 24)	15.55	16.95	16.33
	<i>SD</i>	4.66	3.98	4.32
	CI	[13.89; 17.20]	[15.70; 18.19]	[15.34; 17.33]
Video-based	<i>n</i>	34	34	68
	<i>M</i> (out of 24)	15.82	17.59	16.71
	<i>SD</i>	3.67	3.03	3.46
	CI	[14.54; 17.10]	[16.53; 18.64a]	[15.87; 17.54]
Total	<i>n</i>	67	76	143
	<i>M</i> (out of 24)	15.69	17.24	16.51
	<i>SD</i>	4.16	3.58	3.93
	CI	[14.67; 16.70]	[16.42; 18.05]	[15.86; 17.16]

Interaction between mode and proficiency was not significant, as indicated by the second ANOVA (see Table 4.43). Delivery mode had no effect on test-takers’ motivation

perceptions, with a negligible effect size ($\eta^2 < .01$). The main effect of proficiency was significant. Higher-level test-takers were generally more motivated than lower-level test-takers. The size of this effect was small ($\eta^2 = .04$).

Table 4.43

Two-Way Factorial ANOVA on Test-Takers' Motivation Perceptions

Source	<i>df</i>	<i>F</i>	<i>p</i>	η^2
Delivery mode	1	0.49	.485	< .01
Proficiency	1	5.92	.016*	.04
Mode*Proficiency	1	0.08	.784	< .01
Error	139			

Note: *df* – degrees of freedom; *F* – F-statistic; η^2 – partial eta squared; *significant at $\alpha = .05$; *N* = 143

The second ANOVA results did not support the researcher’s hypothesis that test-takers would find the video-based mode more interesting than the audio-only mode, irrespective of proficiency. Regardless of mode, higher-proficiency learners were more motivated than lower-proficiency learners.

Authenticity perceptions. The audio-only mode yielded authenticity ratings of a slightly higher magnitude, $M = 19.04$, $SD = 3.25$ (audio-only) and $M = 18.26$, $SD = 3.38$ (video-based), as indicated by the total scores for mode (Table 4.44, last column). Being greater than the midpoint of $24/2 = 12$, the means also showed that both modes were considered authentic, though to slightly different degrees. Within each delivery mode, lower-level test-takers gave slightly lower authenticity ratings than higher-level test-takers did. Total scores for the lower and higher proficiency categories also reflected this trend ($M = 18.28$ out of 24, $SD = 3.48$ and $M = 19.01$, $SD = 3.16$ respectively).

Similar to the insignificant interaction term, delivery mode was found to have no effect on test-takers’ authenticity perceptions, as shown in Table 4.45. The size of this effect was trivial ($\eta^2 = .01$). Similarly, the main effect of proficiency was not significant

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

with a small effect size ($\eta^2 = .01$). Referring back to Table 4.44, we see that seeming differences in ratings by mode did not reach statistical significance.

Table 4.44

Descriptive Statistics for Test-Takers' Authenticity Perceptions

Delivery mode		Proficiency		Total
		Lower	Higher	
Audio-only	<i>n</i>	33	42	75
	<i>M</i> (out of 24)	18.48	19.48	19.04
	<i>SD</i>	3.17	3.29	3.25
	CI	[17.36; 19.61]	[18.45; 20.50]	[18.29; 19.79]
Video-based	<i>n</i>	34	34	68
	<i>M</i> (out of 24)	18.09	18.44	18.26
	<i>SD</i>	3.79	2.96	3.38
	CI	[16.76; 19.41]	[17.41; 19.47]	[17.45; 19.08]
Total	<i>n</i>	67	76	143
	<i>M</i> (out of 24)	18.28	19.01	18.67
	<i>SD</i>	3.48	3.16	3.33
	CI	[17.43; 19.13]	[18.29; 19.74]	[18.12; 19.22]

Table 4.45

Two-Way Factorial ANOVA on Test-Takers' Authenticity Perceptions

Source	<i>df</i>	<i>F</i>	<i>p</i>	η^2
Delivery mode	1	1.65	.201	.01
Proficiency	1	1.46	.230	.01
Mode*Proficiency	1	0.33	.568	< .01
Error	139			

Note: *df* – degrees of freedom; *F* – F-statistic; η^2 – partial eta squared; *significant at $\alpha = .05$; *N* = 143

The third ANOVA results did not support the researcher’s hypothesis that participants would find the video-based mode more authentic than the audio-only mode, regardless of proficiency. Test-takers thought the test was realistic regardless of mode and proficiency.

Video use perceptions. For video use, test-takers expressed their agreement about whether listening tests should have videos. The audio-only and video-based modes yielded similar video use ratings, as indicated in Table 4.46 ($M = 12.56$ out of 18, $SD =$

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

2.40 for audio-only and $M = 12.59$, $SD = 2.48$ for video-based). Greater than the midpoint of $18/2 = 9$, these means also showed that test-takers' perceptions within both modes supported the use of videos in listening tests. Within each delivery mode, lower-level test-takers gave slightly higher video use ratings than higher-level test-takers did. Total scores for the lower and higher proficiency categories also reflected this trend ($M = 13.06$ out of 18, $SD = 2.80$ and $M = 12.14$, $SD = 1.98$ respectively).

Table 4.46

Descriptive Statistics for Test-Takers' Video Use Perceptions

Delivery mode		Proficiency		Total
		Lower	Higher	
Audio-only	<i>n</i>	33	42	75
	<i>M</i> (out of 18)	13.24	12.02	12.56
	<i>SD</i>	2.63	2.07	2.40
	CI	[12.31; 14.18]	[11.38; 12.67]	[12.01; 13.11]
Video-based	<i>n</i>	34	34	68
	<i>M</i> (out of 18)	12.88	12.29	12.59
	<i>SD</i>	2.98	1.83	2.48
	CI	[11.84; 13.92]	[11.65; 12.93]	[11.99; 13.19]
Total	<i>n</i>	67	76	143
	<i>M</i> (out of 18)	13.06	12.14	12.57
	<i>SD</i>	2.80	1.98	2.43
	CI	[12.38; 13.74]	[11.70; 12.60]	[12.17; 12.97]

Delivery mode did not have an effect on test-takers' video use perceptions, as shown in Table 4.47. The effect size was trivial ($\eta^2 < .01$). The main effect of proficiency was significant with a small effect size ($\eta^2 = .04$). We see that lower-level test-takers supported the use of videos to a significantly greater extent than higher-level test-takers (see Table 4.46).

The fourth ANOVA results did not support the researcher's hypothesis that the video-based group of test-takers would be more in favor of using content-rich videos in L2 academic tests than the audio-only group, with no regard to proficiency.

Table 4.47

Two-Way Factorial ANOVA on Test-Takers' Video Use Perceptions

Source	<i>df</i>	<i>F</i>	<i>p</i>	η^2
Delivery mode	1	0.01	.910	< .01
Proficiency	1	5.01	.027*	.04
Mode*Proficiency	1	0.61	.436	< .01
Error	139			

Note: *df* – degrees of freedom; *F* – F-statistic; η^2 – partial eta squared; *significant at $\alpha = .05$; *N* = 143

Summary. The results for research question 2.1 provided limited evidence for the researcher’s overarching hypotheses that test-takers would lend support for using content-rich videos in L2 academic listening tests. Difficulty perceptions were significantly affected by mode, showing that the video-based mode was perceived to be easier than the audio-only mode regardless of proficiency. However, test-takers’ perceptions on motivation, authenticity, and video use were not affected by delivery mode.

Apart from comparing perceptions by mode, video use ratings from the video-based group of test-takers were valuable per se, because they showed to what extent test-takers would like to see content-rich videos (not just any videos) in listening tests. The video-based group and the audio-only group favored the inclusion of content-rich videos in tests, regardless proficiency level.

Research question 2.2. Research question 2.2 examined teachers’ perceptions in the same four areas as research question 2.1 did on test-takers’ perceptions, namely difficulty, motivation, authenticity, and video use. Specifically, it asked: How does teachers’ background (i.e., geographical region, education, and L2 teaching-related experience) affect their perceptions about the effect of content-rich videos on listening difficulty, motivation, and authenticity, and use of content-rich videos in tests? It was hypothesized that teachers’ background would have no effect on their perceptions. This

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

section reports on the results of testing these hypotheses using the data from 310 ESL/EFL teachers (note: samples sizes for each perception construct were slightly different due to different outlier patterns).

Operationalization of the variables. There were four dependent variables, namely difficulty, motivation, authenticity, and video use. Each of them was measured using four questions on an 1-to-6 ordinal scale each (1 – strongly disagree, 6 – strongly agree). For example, difficulty perceptions were estimated with the four questions eliciting teachers' agreement that a lecture excerpt in the video-based mode was easier than the same excerpt with the video (see Appendix J). The answers on these four questions were combined, generating an overall difficulty score ranging from 1 to 24. This composite score was now considered to be on a continuous scale. Analogous procedures were applied to the perception constructs of motivation, authenticity, and video use.

Higher score on any construct indicated stronger support for content-rich videos as facilitators of academic listening comprehension, motivation, authenticity, or language tests. Scores around 16 would indicate that teachers generally opted for 4 (somewhat agree). Scores around 20 would show that teachers generally picked 5 (agree). Scores around 24 would show that teachers chose 6 (strongly agree) on average.

The use of composite scores was supported theoretically and statistically. In theoretical terms, the four questions within each of the dependent variables of difficulty, motivation, authenticity, and video use were designed to measure the same corresponding construct. From there, the reliability analyses supported the use of composite scores statistically. Table 4.48 depicts Cronbach's alpha indices by construct, along with the number of items (k), score ranges, and item-total correlation ranges. Adequate reliability

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

indices (> .70) and item-total correlations (> .30) justified the use of composite scores as measures of the four constructs (see Table 4.48).

Table 4.48

Internal Consistency Reliabilities for Teachers' Perceptions

DV	<i>k</i>	Score range	<i>α</i>	Item-total correlation range
Difficulty	4	1-24	.70	.41-.63
Motivation	4	1-24	.79	.48-.73
Authenticity	4	1-24	.77	.43-.64
Video use	4	1-24	.88	.72-.79

There were three independent variables, namely geographic region, education level, and professional experience (henceforth, region, education, and experience respectively). The independent variables were operationalized as reflected in Table 4.49. Each of the three independent variables was measured with one item in the demographics section of the teachers' questionnaire (see Appendix J) and ranged on a scale of nominal and ordinal values from one to four or one to five.

Table 4.49

Operationalizations of the Independent Variables for RQ 2.2

Independent variable	<i>k</i>	Score range	Values
Geographic region	1	1-5	1 – Asia and Oceania 2 – Europe and Eurasia 3 – Caribbean, Central, and South America 4 – Africa and the Middle East 5 – North America
Education level	1	1-4	1 – Certificate 2 – Bachelor's 3 – Master's 4 – Doctorate
Professional experience	1	1-5	1 – 1 to 5 years 2 – 6 to 10 years 3 – 11 to 15 years 4 – 16 to 20 years 5 – more than 20 years

Note: *k* = number of items

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Assumption check. Four three-way ANOVAs were run to examine the effects of geographic region, level of education, and amount of experience on teachers' perceptions of listening difficulty, motivation, authenticity, and use of content-rich videos in academic listening tests.

For each ANOVA, the following assumptions were checked: (a) independence of observations, (b) no significant outliers, (c) normality of the dependent variables' distributions for each combination of the groups of the independent variables, and (d) homogeneity of variance for each combination of the groups of the independent variables. Independence of observations was warranted by the design of the study. Several significant outliers ($z \geq 3.29$) were found, including four for difficulty, three for motivation, one for authenticity, and three outliers for video use. These outliers were treated as missing values in the ANOVA analyses.

The normality assumption was generally not met because the data on each dependent variable were negatively skewed. Because ANOVA is known to be robust against violations of normality, this violation was not considered critical. The assumption of the equality of error variances was checked by running four Levene's tests for each of the four constructs. None of the four test statistics reached significance, thereby upholding the equality-of-variance assumption.

Difficulty perceptions. Teachers perceived videos as strong facilitators of test-takers' academic listening comprehension (e.g., "The video helps learners understand what they hear."), as shown by the total mean in Table 4.50 ($M = 20.89$ out of 24, $SD = 2.67$). The table shows the following information for each value of the three independent variables: the sample size, mean, standard deviation, confidence interval, and the median.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Region-wise, teachers in Caribbean, Central, and South America seemed to consider videos slightly more helpful than teachers in the other four geographic locations.

However, this observation may be misleading because it relies on the trivial sample size ($n = 7$). Teachers at different education and experience levels had comparable difficulty perceptions, with the means slightly higher than 20.

Table 4.50

Descriptive Statistics for Teachers' Perceptions on Difficulty

Independent variable	Value	<i>n</i>	<i>M</i>	<i>SD</i>	CI	Median
Region	Asia and Oceania	110	20.95	2.71	[20.44; 21.46]	21.00
	Europe and Eurasia	36	20.47	2.89	[19.49; 21.45]	20.00
	Caribbean, Central, and South America	7	22.00	2.45	[19.73; 24.27]	23.00
	Africa and the Middle East	50	20.62	2.60	[19.88; 21.34]	21.00
	North America	97	21.09	2.63	[20.56; 21.62]	21.00
Education	Certificate	24	20.33	3.29	[18.94; 21.72]	20.00
	Bachelor's	45	20.96	2.49	[20.21; 21.70]	21.00
	Master's	183	20.85	2.60	[20.47; 21.23]	21.00
	Doctorate	48	21.38	2.83	[20.56; 22.20]	22.00
Experience	1 to 5 years	49	20.47	2.58	[19.73; 21.21]	21.00
	6 to 10 years	53	20.83	2.80	[20.06; 21.60]	22.00
	11 to 15 years	45	21.20	2.46	[20.46; 21.94]	22.00
	16 to 20 years	59	21.30	2.76	[20.57; 22.00]	21.00
	more than 20 years	94	20.80	2.71	[20.24; 21.35]	21.00
Total	-	306	20.89	2.67	[20.59; 21.19]	21.00

Teachers' perceptions about listening difficulty were not affected by geographic region, education level, or the amount of teaching-related experience, as shown in Table 4.51. Having discounted the four insignificant interaction terms, none of the three main effects were found significant ($p < .05$). The effect sizes for the three main effects were negligible (i.e., $\eta^2 = .01$).

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

The results of the first ANOVA supported the researcher’s hypothesis that teachers would consider content-rich videos helpful for listening comprehension, regardless of their professional location, education, or experience.

Table 4.51

Three-Way Factorial ANOVA on Teachers’ Perceptions on Difficulty

Source	<i>df</i>	<i>F</i>	<i>p</i>	η^2
Region	4	0.62	.652	.01
Education	3	0.65	.581	.01
Experience	4	0.70	.592	.01
Region*Education	10	1.05	.406	.04
Region*Experience	14	1.09	.368	.06
Education*Experience	12	1.35	.194	.07
Region*Education*Experience	18	1.27	.211	.09
Error	233			

Note: *df* – degrees of freedom; *F* – F-statistic; η^2 – partial eta squared; *N* = 306

Motivation perceptions. Teachers perceived videos as strong motivators for test-takers’ academic listening comprehension (e.g., “The video makes this listening more engaging for students.”), as indicated by the total mean in Table 4.52 below ($M = 20.62$, $SD = 2.89$). Region-wise, teachers in Caribbean, Central, and South America seemed to consider videos slightly more motivating than teachers in the other four geographic locations (however, $n = 7$). Teachers at different education and experience levels had comparable motivation perceptions, with the means slightly higher than 20 across the levels. Note that sample sizes in Tables 4.50 and 4.52 are not identical due to different outlier patterns in the difficulty and motivation data.

Region, education and experience did not play a role in teachers’ perceptions about motivation, as shown in Table 4.53 and Appendix L. There was one significant interaction effect between education and experience, $F(12, 232) = 1.90$, $p = .035 < .05$, $\eta^2 = .09$. However, Bonferroni-corrected *post hoc* analyses for this interaction did not

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

reveal any significant pairwise differences (see Appendix L). Region was found to have no effect on teachers’ motivation perceptions, $F(4, 232) = 0.65, p = .628 > .05, \eta^2 = .01$.

Table 4.52

Descriptive Statistics for Teachers’ Perceptions on Motivation

Independent variable	Value	<i>n</i>	<i>M</i>	<i>SD</i>	CI	Median
Region	Asia and Oceania	112	21.11	2.78	[20.59; 21.63]	22.00
	Europe and Eurasia	36	20.81	2.97	[19.80; 21.81]	22.00
	Caribbean, Central, and South America	7	22.14	2.04	[20.26; 24.03]	23.00
	Africa and the Middle East	49	21.08	2.82	[20.27; 21.89]	22.00
	North America	97	20.84	2.95	[20.24; 21.43]	21.00
Education	Certificate	24	20.00	3.28	[18.61; 21.34]	20.00
	Bachelor’s	44	21.23	2.56	[20.45; 22.01]	21.50
	Master’s	185	21.01	2.81	[20.60; 21.41]	22.00
	Doctorate	48	21.29	2.95	[20.44; 22.15]	22.00
Experience	1 to 5 years	49	20.89	2.81	[20.05; 21.66]	21.00
	6 to 10 years	52	21.12	2.92	[20.30; 21.93]	21.00
	11 to 15 years	47	21.09	2.82	[20.26; 21.91]	22.00
	16 to 20 years	60	21.08	2.96	[20.32; 21.85]	22.00
	more than 20 years	93	20.93	2.80	[20.35; 21.50]	22.00
Total	-	307	20.62	2.89	[20.62; 21.27]	22.00

Table 4.53

Three-Way Factorial ANOVA on Teachers’ Perceptions on Motivation

Source	<i>df</i>	<i>F</i>	<i>p</i>	η^2
Region	4	0.65	.628	.01
Education	3	2.14	.096	.03
Experience	4	1.19	.314	.02
Region*Education	10	0.70	.724	.03
Region*Experience	14	0.72	.751	.04
Education*Experience	12	1.90	.035*	.09
Region*Education*Experience	18	0.84	.648	.06
Error	234			

df – degrees of freedom; *F* – F-statistic; η^2 – partial eta squared; *significant at $\alpha = .05$; *N* = 307

The results of the second ANOVA supported the researcher’s hypothesis that teachers would find content-rich videos to be motivating, regardless of their professional location, education, or experience.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Authenticity perceptions. Teachers perceived videos as authentic on average (e.g., “This video makes this listening more realistic.”), as indicated by the grand mean in Table 4.54 ($M = 19.72$, $SD = 3.30$). Teachers at different education and experience levels had comparable authenticity perceptions, with the means of about 19-20 across the levels.

Table 4.54

Descriptive Statistics for Teachers’ Perceptions on Authenticity

Independent variable	Value	<i>n</i>	<i>M</i>	<i>SD</i>	CI	Median
Region	Asia and Oceania	112	19.80	3.60	[19.12; 20.47]	20.00
	Europe and Eurasia	35	19.49	3.24	[18.37; 20.60]	20.00
	Caribbean, Central, and South America	7	21.00	2.58	[18.61; 23.39]	22.00
	Africa and the Middle East	51	19.59	3.41	[18.63; 20.55]	20.00
	North America	98	19.74	3.06	[19.12; 20.35]	20.00
Education	Certificate	23	19.22	3.49	[17.71; 20.73]	19.00
	Bachelor’s	46	20.13	3.36	[19.13; 21.13]	21.00
	Master’s	186	19.56	3.34	[19.09; 20.06]	20.00
	Doctorate	48	20.21	3.16	[19.29; 21.13]	21.00
Experience	1 to 5 years	49	20.59	2.49	[19.88; 21.31]	21.00
	6 to 10 years	53	19.19	3.78	[18.15; 20.23]	20.00
	11 to 15 years	47	19.34	3.48	[18.32; 20.36]	20.00
	16 to 20 years	60	20.27	3.28	[19.42; 21.11]	21.00
	more than 20 years	94	19.45	3.32	[18.77; 20.13]	20.00
Total	-	309	19.72	3.30	[19.35; 20.09]	20.00

Teachers’ perceptions on listening authenticity were not affected by geographic region, education level, or the amount of teaching-related experience, as reflected in Table 4.55 below. Similar to the four insignificant interaction terms, none of the three main effects were found significant ($p < .05$). The effect sizes for the three main effects were small ($\eta^2 = .01-.03$).

The results of the third ANOVA supported the researcher’s hypothesis that teachers would consider content-rich videos authentic, regardless of professional location, education, or experience.

Table 4.55

Three-Way Factorial ANOVA on Teachers' Perceptions on Authenticity

Source	<i>df</i>	<i>F</i>	<i>p</i>	η^2
Region	4	0.22	.929	< .01
Education	3	1.27	.287	.02
Experience	4	1.51	.201	.03
Region*Education	10	1.33	.217	.05
Region*Experience	14	0.81	.663	.05
Education*Experience	12	1.00	.446	.05
Region*Education*Experience	17	1.27	.210	.08
Error	237			

df – degrees of freedom; *F* – F-statistic; η^2 – partial eta squared; *N* = 143

Video use perceptions. Teachers perceived videos as desired complements to large-scale second language academic listening tests (e.g., “Large-scale academic English listening tests should have videos like this.”), as shown in Table 4.56 (*M* = 19.12, *SD* = 4.06). Region-wise, teachers seemed to have equally favorable opinions about using videos in listening tests. Similarly, teachers at different education and experience levels had comparable video use perceptions, with the means of about 19-20 across the levels.

Table 4.56

Descriptive Statistics for Teachers' Perceptions on Video Use

Independent variable	Value	<i>n</i>	<i>M</i>	<i>SD</i>	CI	Median
Region	Asia and Oceania	112	19.21	4.26	[18.41; 20.01]	20.00
	Europe and Eurasia	32	18.92	3.79	[17.59; 20.23]	19.00
	Caribbean, Central, and South America	7	18.43	5.06	[13.75; 23.11]	18.00
	Africa and the Middle East	51	19.02	4.36	[17.79; 20.25]	20.00
	North America	97	19.35	3.76	[18.59; 20.11]	20.00
Education	Certificate	23	18.83	3.76	[17.20; 20.45]	20.00
	Bachelor's	46	19.78	3.89	[18.63; 20.94]	21.00
	Master's	184	18.92	4.13	[18.32; 19.52]	20.00
	Doctorate	48	19.73	4.14	[18.53; 20.93]	20.00
Experience	1 to 5 years	49	19.84	3.03	[18.97; 20.71]	21.00
	6 to 10 years	52	18.84	4.28	[17.65; 20.04]	20.00
	11 to 15 years	47	19.53	4.08	[18.34; 20.73]	20.00
	16 to 20 years	60	19.33	4.25	[18.24; 20.43]	20.00
	more than 20 years	93	18.72	4.30	[17.83; 19.61]	19.00
Total	-	307	19.12	4.06	[18.66; 19.57]	20.00

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Teachers' perceptions about video use in listening tests were not affected by geographic region, education level, or the amount of teaching-related experience, as indicated in Table 4.57 below. Similar to the four insignificant interaction terms, none of the three main effects were found significant ($p < .05$). The effect sizes for the three main effects were very small (η^2 of .01 to .02).

Table 4.57

Three-Way Factorial ANOVA on Teachers' Perceptions on Video Use

Source	<i>df</i>	<i>F</i>	<i>p</i>	η^2
Region	4	0.66	.619	.01
Education	3	1.11	.348	.01
Experience	4	1.01	.403	.02
Region*Education	10	1.09	.368	.04
Region*Experience	14	0.78	.687	.05
Education*Experience	12	0.82	.634	.04
Region*Education*Experience	17	0.97	.488	.07
Error	235			

df – degrees of freedom; *F* – F-statistic; η^2 – partial eta squared; *N* = 307

The results of the fourth ANOVA supported the researcher's hypothesis that teachers would be in favor of including content-rich videos in standardized L2 academic listening tests, regardless of professional location, education, or experience.

Summary. Regardless of professional location, education, and experience, teachers agreed that content-rich videos decreased listening difficulty, increased motivation towards listening, improved authenticity of listening comprehension, and should be included in large-scale second language academic listening tests. It is worth noting that teachers' favorable opinions about the videos were somewhat stronger for difficulty and motivation and weaker for authenticity and video use in large-scale listening tests. However, this conclusion should be verified in follow-up studies.

Chapter 5

Discussion

This dissertation study examined the role of content-rich videos in second language (L2) tests that measure listening comprehension of academic lectures. It followed the argument-based validity approach to justify the use of video-inclusive L2 academic listening assessment constructs (Kane, 2004; 2006; 2013; Chapelle et al., 2008). The argument-based validity approach requires test score interpretations and uses to be based on well-defined and empirically supported inferences. There are six inferences in the most recent validity framework, including test domain (domain definition), evaluation, generalization, explanation, extrapolation, and utilization. Each inference is backed by theoretical and/or empirical evidence. Out of these six inferences, the main focus of the study was the explanation inference, which deals specifically with the nature of the measured assessment construct.

The explanation inference warrants that the test scores are attributed to the construct. Properly speaking, the variation among scores should only be due to different degrees to which test-takers possess the measured construct, which was test-takers' academic English listening comprehension in this study. Thus, differences in scores from equivalent-ability student groups is reflective of differences in the constructs the groups were measured on. If theory-informed, this difference may be used as divergent evidence supporting the defined construct (Chapelle et al., 2008). This dissertation study examined differences in test-takers' performances on the video-based versus the audio-only versions of the developed academic listening comprehension (ALC) test. The video-based ALC test represented a theory-informed construct that included videos displaying

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

content-rich visual information pertinent to academic contexts, such as graphs and pictures. The audio-only ALC test represented what was considered a deficient audio-only construct. The following two types of backings were sought for the explanation inference: (1) test and item difficulties were affected by content-rich visuals as a theory-informed construct-relevant factor, and (2) the visually content-rich L2 academic listening construct was supported by test stakeholders, namely test-takers and teachers.

This chapter is centered on these two types of explanation inference backings. It discusses the findings related to each backing, followed by implications of the findings, limitations of the study, and recommendations for future research. The chapter is bracketed by the summary of findings and the conclusion.

Summary of Findings

This sections summarizes the findings through the lens of explanation inference backings relating to both ALC test difficulty and stakeholders' perceptions.

Test difficulty. The first backing for the explanation inference was based on evidence from answering the first research question in the study: *Do content-rich videos affect L2 academic listening comprehension difficulty?* The findings suggested a positive answer to this question, as shown below.

(a) Videos made the overall ALC test easier for test-takers, with no moderating effect of proficiency or item type.

(b) Videos made video-dependent ALC items easier for test-takers. Proficiency and item type did not contribute to this effect. At the individual item level, about 70% of video-dependent items were easier in the video-based mode than in the audio-only mode, although nearly 30% of video-dependent items were harder with videos.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

(c) Videos made video-independent ALC items harder for test-takers in general. Proficiency and item type accounted for this effect. For lower-proficiency test-takers, video-independent global items were harder with videos than with audio-only. At the item level, most of video-independent items were harder in the video-based mode (90%), though some video-independent items were easier with videos than without (10%).

(d) There was a weak positive relationship between self-reported viewing behavior rates and ALC total scores in general. This relationship was stronger for lower-level students but was almost absent for higher-level students.

Stakeholders' perceptions. The second piece of evidence backing the explanation inference stemmed from answering the second research question: *Do stakeholders' perceptions lend support for using content-rich videos in the L2 academic listening assessment construct?* A positive answer for this question was largely attained, as suggested by the findings below.

(a) Test-takers found the video-based mode easier than the audio-only mode, regardless of proficiency level. Additionally, test-takers were in favor of including content-rich videos in L2 academic listening tests. Perceptions about motivation towards and authenticity of listening were equally favorable across the two modes, regardless of listening proficiency.

(b) Teachers found the video-based mode to be facilitating and motivating, and to increase authenticity of academic listening comprehension. They were also in favor of using content-rich videos in tests. These perceptions were independent of teachers' professional location, level of relevant education, and amount of L2 teaching experience.

Listening Comprehension Difficulty

Previous research and theory suggested that content-related visual information generally decreases listening comprehension difficulty (Rost, 2016). There was also a theory-driven expectation that lower-level learners' comprehension may be adversely affected by the presence of visuals (Mayer, 2005; Paivio, 1991; 2006). The present study examined the effects of content-rich videos on L2 listening comprehension, which was measured by both video-dependent and video-independent items. It also investigated the role of proficiency and item type in this effect.

This section first discusses how difficulty of video-dependent items compared in the audio-only and the video-based modes. Then, it expands on difficulty of video-independent items across the two delivery modes. Next, the findings regarding viewing behavior are discussed, followed by a brief summary.

Items with video-dependent design. Results showed that video-dependent items were easier with content-rich videos than with audio-only. This indicated that, overall, test-takers understood lecture points better if these points were explained, illustrated, and/or organized in the video than if no videos were given. Both lower- and higher-proficiency test-takers were favored in this mode of presentation, on both local and global lecture items. These results echo findings of other studies that found facilitative effects of videos on listening comprehension (e.g., Baltova, 1994; Lee & Lee, 2015; Lesnov, 2017; Shin, 1998). Videos in these studies contained some content-related visual information. In Lesnov's study, the video-based academic testlet with the highest amount of content-related visuals and the highest number of video-dependent items was easier than the audio-only version of the same testlet, with no such effects for other testlets. Findings

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

from this and the present study suggest that if the video is sufficiently rich in content-related visuals, it becomes capable of cueing lecture points. If cued lecture points are tested, the presence of content-rich videos may facilitate test-takers academic listening comprehension. Contemporary theory upholds this conclusion (Rost, 2016; Ur, 1984).

Improved academic listening comprehension on video-dependent items in the video-based mode was not affected by test-takers' proficiency. Content-rich videos did not have reverse effects for lower- and higher-level learners, as hypothesized (Mayer, 2005; Paivio, 1991, 2006). It may mean that content-rich videos were equally helpful for academic listeners of low-intermediate and high-intermediate abilities. This was dissimilar to the findings in Lesnov (2017), where higher-intermediate students were benefited by content-related visuals but lower-intermediate students were not affected. However, lower- and higher-level students in Lesnov's study watched videos with significantly lower amounts of content-related visuals. Other studies either used context videos or left out proficiency as a variable.

One possible reason for video-dependent items being easier with videos than with audio-only for both proficiency levels is that, perhaps, proficiency of lower test-takers was not low enough to cause confusion in the dual mode. While significantly lower relative to the higher-proficiency learners, the lower-proficiency learners' linguistic capacity might have still been sufficient for successfully processing both oral and visual modes at once, without serious ramifications for comprehension. Also, the lower-versus-higher proficiency dichotomy may not have been adequate for detecting video effects at far ends of proficiency. For instance, video effects on low beginners could be washed away in the overarching lower-proficiency category. Future studies should investigate

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

effects of mode on comprehension at different proficiency levels in a more fine-grained manner. It was hard to do in the present study due to the anchor test's limited capacity to discriminate between test-takers' abilities beyond the two-level dichotomy.

There might be another explanation for finding no moderating effect of listening proficiency on the relationship between delivery mode and video-dependent items' difficulty. It may have been the case that the hypotheses in the present study were not properly aligned with the issue of item video-dependence. Following Gruba (2004), Paivio & Lambert (1981), and Mayer (2005; 2009), it was hypothesized that video-dependent items would be easier for higher-level but harder for lower-level test-takers due to the latter group's limited linguistic ability. In other words, it was expected that lower-level learners would be overwhelmed by the need to process both the auditory channel and the visual channel. However, many video-dependent items were clearly cued by content-related visuals in addition to language, meaning that many of them could be answered based on the video input alone, as suggested by the results of the muted version of the ALC test in this study. This may have largely reduced the processing load for lower-level students in the video-based mode, which could have explained the discovered positive effect of content-rich videos. Following this line of reasoning, it may have been more reasonable to hypothesize that video-dependent items would be easier with video than with just audio, regardless of listening proficiency.

The results of item-level analyses revealed that most video-dependent items were easier with content-rich videos, regardless of proficiency. This further corroborates the above-stated findings and agrees with previous research (i.e., Wagner, 2010b). Similar to

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

this study, Wagner found that some items cued by video-based pictures were easier in the video mode than in the audio-only mode.

Four out of the 14 video-dependent items were unexpectedly harder in the video-based mode than in the audio-only mode. Although designed to be video-dependent, these items seemed to have not been cued by content-rich videos. Rather, many test-takers were likely misguided by videos while answering these questions. While it was impossible to objectively determine the reasons for this outcome, one conjecture could be made. Two of the harder-with-video video-dependent items were global main ideas. Notably, all the four main-idea items in the test (two video-dependent and two video-independent main-idea items) were harder in the video-based mode than in the audio-only mode regardless of proficiency. This may suggest that main-idea items were especially challenging for learners, and this challenge increased in the presence of content-rich videos, even for video-dependent items.

The remaining video-dependent items that were harder with videos than with just audio were detail-oriented. Reasons behind this finding were hard to determine based solely on quantitative data. Further qualitative process-oriented analyses of interviews with test-takers could shed more light on this unexpected outcome (Gruba, 2014; Suvorov, 2013, 2015a). As a possible explanation, both items had one distractor that could potentially be triggered by videos. Videos may have had graphs that somehow related to the distractor in terms of shape, form, or position. This link may have looked stronger to test-takers than the link between the graph and the key, especially for test-takers who missed the point tested by the item and were using whatever video fragments flashed out in their memories. As a result, some test-takers may have been prone to pick

the distractor. This speaks to the importance of thoroughly screening video-dependent items for unwanted tricky connections to their distractors, which can be achieved by piloting the test with and without videos and carrying out distractor analyses by mode.

Items with video-independent design. Next, 10 video-independent items were generally harder in the video-based mode than in the audio-only mode. This effect was particularly strong on global video-independent items for lower-level learners. This means that lower-proficiency learners understood global lecture points worse if they were not explained, illustrated, and/or organized in the video but the video was still present. In other words, test-takers comprehended a no-video lecture better than a lecture with a video that did not cue test questions. This is similar to findings in Suvorov's (2009) and Pusey and Lenz's (2014) studies. Both studies found that academic English tests were harder with the video than in the audio-only condition for high- and intermediate-proficiency students respectively. However, both studies worked with context videos, which, by definition, did not have content-related cues (Bejar et. al, 2000, Ginther, 2002).

This negative outcome can perhaps be linked to lower-level learners' linguistic and cognitive capacities. Also, the visual channel requires additional visual processing (e.g., Mayer, 2005; 2009; Paivio, 1979; 1991; 2006). Given that lecture stimuli already impose a burden on test-takers (Lynch, 2011), an additional load from the visual channel may have cognitively overwhelmed lower-level L2 learners, negatively affecting their comprehension of video-uncued lecture points. Because, by design, visual information was weakly related to the content of video-independent items, it might have been a distraction. Further, global comprehension is generally harder than local comprehension, and more so for lower-level learners (Becker, 2016; Hansen & Jensen, 1994; Shohamy &

Inbar, 1991). This may also have contributed to the significance of the negative video effect on lower-level test-takers' understanding of video-uncued global lecture points.

Reflecting back, the researcher's hypotheses for video-independent items seem to have been erroneous. Expecting video-independent items to be unaffected by mode regardless of proficiency did not reflect the dual processing challenges for lower-level students, especially in the presence of mostly unrelated, potentially distracting content-rich visual information.

Individual items' performance showed that, regardless of item type (i.e., local vs. global), most video-independent items were harder in the video-based mode for lower-level students. This resonates with Batty's study (2015), which found higher difficulty of two video-uncued items in the video-based mode than in the audio mode. Batty suggested that test-takers were confused by the speaker's facial expression for one item while remaining at a loss to explain the effect on the other item. In the present study, no potentially confusing facial expressions or gestures were identified. Rather, mere presence of visual information is believed to have been confusing for lower-level test-takers due to its potential to congest their low-capacity speech processing.

Overall test difficulty. Recall that four out of the 14 video-dependent items were unexpectedly harder with videos than with audio-only while three out of the 10 video-independent items were unexpectedly easier with videos. Collectively, there were two more items that were easier with videos ($k = 13$) than items that were harder with videos ($k = 11$). One might use this to explain why the overall ALC test was easier with videos than with audio-only. There is another possible interpretation of this finding, however. Despite having seven video-independent and even four video-dependent items that were

harder with video than with audio-only, the video-based test was still easier on the whole. This means that negative videos effects on items did not cancel out positive video effects on items. It may show that the facilitative power of the video effect was considerably larger than its detrimental impact. This strengthens the overall finding of content-rich videos' facilitative effect on academic listening comprehension. Either of these interpretations suggests that previous studies that did not find the effect of content-related visuals may have been under-researched (e.g., Baltova, 1994; Lesnov, 2017; Suvorov, 2013; 2015b). Some items in these studies may have been facilitated but others adversely affected by videos, which may have cancelled out the overall effect of videos. Though a speculation, it is a possibility showing the importance of item-level analysis and controlling for item video-dependence.

Viewing behavior. Test-takers' self-reported viewing behavior ratings further supported the argument for including content-rich visuals in academic listening tests. Overall, viewing ratings showed that test-takers were oriented to videos most of the lecture time, regardless of proficiency. This is in agreement with Wagner's study (2007). Wagner found that test-takers attended to lecture-like videos for up to 67% of the time. Next, weak positive correlations between viewing rates and the ALC test scores did not disprove that, at least for lower-level students, the more attentively test-takers watched the videos, the higher their scores were. Though indirectly, it shows that content-rich videos were somewhat helpful for understanding the lectures. This conflicts with the finding from Wagner's (2010b) and Suvorov's (2015a) studies, which reported a weak negative correlation and no correlation between viewing behavior and test performance

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

respectively. Note that both studies used videos that were considerably less content-oriented relative to the present study.

The fact that the correlation between viewing behavior and comprehension was low does not challenge the video-based L2 academic listening construct. It still shows that viewing behavior and academic listening comprehension are two related constructs. Viewing is just one part of academic listening, in addition to other construct-related facets, such as motivation, memory, sociocultural competence, metacognition, background knowledge, and others (e.g., Buck, 2001; Flowerdew & Miller, 2005; Field, 2008; Rost, 2016; Vandergrift & Goh, 2012). As a segment of the overall construct, viewing behavior should not be expected to correlate strongly with a measure of the construct. A significant moderate correlation in the .03-.50 range might be indicative of a relationship sufficient enough to claim construct relevance.

Stakeholders' Perceptions

The second assumption for the explanation inference was that the video-inclusive construct would be favorably perceived by test stakeholders. To back this assumption, perceptions of test-takers and teachers were elicited.

Test-takers' perceptions. Visuals in general and content-related visuals in particular are often perceived by test-takers as decreasing listening difficulty, and increasing motivation and authenticity (e.g., Ockey, 2007; Suvorov, 2015b; Wagner, 2010a). This study echoed these findings only for listening difficulty. Test-takers that watched the videos found lectures easier than test-takers that had access only to audios. Reflecting theoretical expectations for visual effects, this finding can be used as evidence

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

for including visual information in L2 listening constructs (Gruba, 2014; Flowerdew & Miller, 2005; Rost, 2016).

However, motivation and authenticity perceptions were not affected by the presence of visuals. This finding conflicts with prior research (Cubilo & Winke, 2007; Ockey, 2007; Parry & Meredith, 1984; Wagner, 2010a). One possible reason for this conflict is the design of the studies. Previous studies asked test-takers to compare audio-only and video-based modes either hypothetically, with participants exposed to only one of the modes, or *de facto*, with the same participants being exposed to each of the two modes. This study employed a different method. It compared the perceptions of two different groups of test-takers, with each group exposed to a different mode. This paralleled the design for comparisons of test-takers' performance on the ALC test and relied on the assumption that each group equivalently represented the targeted population of test-takers. This design allowed for more unbiased estimations of test-takers' perceptions due to the absence of contaminating factors, such as the use of different listening stimuli for estimating the perceptions of different modes and carry-over effects, where the perceptions of one mode influence the opinions about the other.

This design may have been responsible for the lack of difference in the authenticity perceptions. Recall that the authenticity item elicited answers for the following question: "How realistic was this lecture?" Because participants were likely used to taking audio-only listening tests, they may have been prone to giving high authenticity scores in the audio-only mode. The same may have been true for the motivation ratings. This limitation may have undermined comparisons of motivation and authenticity perceptions by mode.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Test-takers' video use perceptions indicated that test-takers in the video-based group were in favor of including content-rich videos in L2 academic listening tests. While reflective of trends in some previous research (e.g., Wagner, 2010a), this finding is largely unique as it applies to content-rich videos. The fact that test-takers were exposed to content-rich videos before providing their video use ratings increases the value of these ratings. It shows that test-takers had an ample opportunity to make informed judgments about suitability of content-rich videos for listening tests. These judgments provided support for using content-rich videos in tests.

Teachers' perceptions. Unlike test-takers, L2 teachers had an opportunity to compare the audio-only mode to the video-based mode in terms of difficulty, motivation, authenticity, and video use. Teachers listened to the audio-only excerpt of the Food Tax lecture first, then they watched the video-based version of the same excerpt, followed by the questionnaire. Regardless of professional background, teachers expressed moderate to strong agreement that content-rich videos decreased comprehension difficulty, increased motivation, and improved authenticity of L2 academic listening. In addition, teachers supported the use of content-rich videos in large-scale standardized L2 academic listening tests. Educational level, geographic location, and amount of teaching experience did not factor in these effects.

These findings confirm previous research into the perceptions of L2 *learners* about visuals and difficulty (e.g., Brett, 1997; Ockey, 2007; Sueyoshi & Hardison, 2005; Wagner, 2008; 2010a), visuals and motivation (Ockey, 2007; Progosh, 1996; Suvorov, 2009; Wagner, 2010b), as well as visuals and authenticity (e.g., Cubilo & Winke, 2013), as described on pp. 68-72 in Literature Review. However, ESL and EFL *teachers'*

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

perceptions about the effects of visuals on listening difficulty and motivation in the present study conflict with EFL teachers' opinions in Coniam's study (2001). Teachers in Coniam's study perceived videos to be distracting and ineffective for improving authenticity. Coniam's (2001) study design can partly explain this discrepancy. Some participants in Coniam's study expressed their opinions about the difficulty of a video-enhanced listening test without having an opportunity to compare it with the audio-only version of the same test. This might have skewed the results and limited their generalizability. Though not stated explicitly, the methodological description in Coniam's study implied that videos were of the context type (i.e., a talk show). If context videos do not contain non-verbal cues helping the comprehension, which could be true for Coniam's study, they indeed may turn out to be distracting (Suvorov, 2015b). Nevertheless, such a conclusion seems less likely when judgment concerns content-rich visuals, which was the case in the present study. Aside from Coniam (2001), no studies looked into how the effects of content-related visuals were judged by L2 teachers.

The findings of this study shed new light on how teachers view the L2 academic listening assessment construct. Teachers' support for using visuals upholds the growing tendency to refine listening constructs by including content-related visual information (e.g., Suvorov, 2013; Wagner, 2010a). In Suvorov's study, L2 learners found content visuals helpful for academic listening, which supported the visually inclusive construct. In Wagner's study, test-takers found the audiovisual mode less difficult and more motivating than the audio-only mode, supporting the author's argument for including visuals in L2 listening tests. Both studies reported on test-takers' perceptions of videos with relatively few content-related visuals. Research into teachers' perceptions of

content-rich visual information in assessment contexts has been missing. Coniam (2001) seems to be the only study that researched teachers' perceptions in listening assessment contexts, but it did not use content visuals, as implied by its methodology. Teachers' perceptions in the present study, therefore, has provided new evidence supporting visually-inclusive L2 academic listening constructs while challenging the opposite view found in some scholarly groundwork (e.g., Buck, 2001; Lado, 1961).

The Interpretive Argument

Although this study was primarily concerned with providing backing for the explanation inference in the interpretive argument for including content-rich visuals in the L2 academic listening construct, it built upon three preceding inferences, including test domain, evaluation, and generalization (Chapelle et al., 2008; Kane, 2004; 2013).

Test domain. The domain definition inference focused on the representativeness of the video-based ALC test content and items of the academic listening target language use (TLU) domain (Bachman & Palmer, 2010). Guided by the related literature, the video-based test was designed to elicit the core processes of input decoding, lexical search, syntactic parsing, meaning construction, and discourse construction, using both the auditory and the visual channels (Field, 2013). The test was characterized by high-density listening input and higher speech rates in lectures, which are considered most typical stimuli in academic contexts (Lynch, 2011). Lectures were accompanied by speakers' non-verbal cues and content-rich visual aids, and followed by assessment tasks requiring to infer main ideas, identify details, and make inferences based on the lecture input (Field, 2009; 2011; Powers, 1985; Richards, 1983). These characteristics were reflective of authentic US-based academic contexts.

Evaluation inference. The warrant for the evaluation inference was that observed scores were consistently awarded. It was backed by three types of evidence, namely reliable scoring methods, properly controlled testing conditions, and appropriate psychometric item properties. First, a dichotomous scoring (i.e., 0 or 1) for the ALC test was performed automatically in the online testing system, which reduced the scoring error. The answer key was reviewed by several L2 teachers who were native English speakers as well as non-native English speakers of high proficiency. No inconsistencies were found in the answer key, eliminating the likelihood of miskeyed items.

Second, environmental factors were controlled, reducing the potential to contaminate test performance (Bachman & Palmer, 2010). Although rather limited due to unproctored settings, the control of unwanted environmental factors in the ALC test administrations was aimed for. Prior to test taking, learners had to listen to instructions urging them to attend to the lectures attentively, to remain in a quiet room with minimal distractions, and to avoid pausing the test, reloading web-pages, or leaving the room. Test-takers were able to start the test upon expressing their agreement with these instructions. After the test, test-takers were asked if they encountered any problems with internet connection, video technology, or sound. If any problems emerged, the response was eliminated. These measures helped to eliminate contaminated responses.

One supposed positive aspect of unproctored online test administrations is reduction in test anxiety. Test anxiety is viewed as construct-irrelevant variance and may adversely impact test-takers' performance (Brindley, 1998). Because the test was taken at preferred times and locations, anxiety might have been lower than in formal high-stakes testing situations. Both the scoring method and administration conditions control were

trialed and piloted, with subsequent revisions made. One major revision was removing time constraints for answering individual items to avoid unnecessary pressure.

Third, psychometric qualities of test items were analyzed for their appropriateness for making norm-referenced decisions. Such decisions were predicated upon a sufficient number of items and their discriminative power so that the test could reliably distinguish between test-takers' proficiency levels (Chapelle et al., 2008). The combination of item response theory methods and classical test theory methods were used for item analysis. The Rasch item reliability and infit statistics, and Cronbach's alpha indices indicated adequate discrimination power, which served as a backing for the evaluation inference. Prototyping and pilot studies were conducted prior to the commencement of the study to identify and rework initially problematic items.

Generalization inference. The generalization inference posited that test would generate similar results for a test-taker across measurement events. In other words, test results were assumed to be generalizable across parallel tasks and forms. It was backed by the adequate Rasch person reliability index. Analogously to internal consistency measures, such as Cronbach's alpha, it is a measure of accuracy of proficiency-based discrimination among test-takers ("Reliability and Separation of Measures," 2017). It indicated the video-based ALC test's capacity to consistently distinguish between low and high performers.

In addition, the content characteristics of the test were controlled by test and item specifications. The definition of content-rich videos was provided along with detailed strategies for video design. A test specification had been developed and the item writing

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

techniques were described to ensure the replicability of the ALC test administrations for future measurement events.

Explanation inference. *Listening comprehension difficulty.* The first research question in this study sought to generate discriminant evidence for comprehension difficulty in the video-based mode and the audio-only mode. Content-rich videos were used to shape a sufficient, theory-informed, visually-rich L2 academic construct. The audio-only construct was believed to be deficient. Content-rich videos increased comprehension of visually exposed local and global lecture items, for both lower- and higher-level learners. This conforms to the existing theory that content-related visuals are helpful for listening comprehension (e.g., Rost, 2016; see pp. 35-37 in Literature Review). This evidence suggests that content-related visuals can introduce construct-relevant variance to test scores because they (a) reflect authentic academic settings better, and (b) affect comprehension in keeping with theoretical expectations (Chapelle et al., 2008; Morell, Garcia, & Sanchez, 2008; Field, 2009; 2011; Lynch, 2011; Richards, 1983; Rost, 2016; Ur, 1984). This conclusion was also supported by the convergent correlational evidence suggesting that there was a positive relationship between viewing behavior rates and listening comprehension. Altogether, these results empirically backed the overall explanation inference stating that expected scores reflected the more precisely defined visual-inclusive construct.

One can assume that the finding of content-rich videos adversely affecting low-level comprehension of lecture points that were not visually covered does not favor the video-inclusive argument. The reason behind this assumption might be that content-rich videos may introduce unfairness for lower-level students on video-independent items in

the test. According to this viewpoint, this fairness breach should be avoided, and possible ways of doing this are discussed in one of the following sections. Another point of view is that such an outcome represents what actually happens in authentic target language use situations, and thus, should not be a matter of concern for test developers. This perspective supports the interpretive argument.

Stakeholders' perceptions. The findings for research question 2 advanced the interpretive argument for using the visually content-rich L2 academic listening construct. Both test-takers' and teachers' opinions indicated that content-rich videos made academic listening comprehension easier. This reinforced the more objective findings of the present study, which showed lower comprehension difficulty in the video mode on video-dependent items. While no increases in listening motivation and authenticity were found based on test-takers' perceptions, teachers' perceptions indicated strong potential of content-rich videos to increase listening motivation and improve listening test authenticity. Finally, both test-takers and teachers considered content-rich videos desirable additions to L2 listening tests. Both difficulty- and perception-related evidence supports the explanation inference, upholding the video-based construct as more valid, compared to traditional visual-free listening constructs.

Summary. The proposed interpretive argument for the inclusion of content-related visual information into the assessment construct of L2 academic listening proficiency have accumulated evidence for the domain definition inference, evaluation inference, generalization inference, and explanation inference, with the explanation inferences playing the key role in the argument. Table 5.1 below summarizes the

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

interpretive argument with regards to warrants, assumptions, backings, and specific analyses used to generate backings for each of the four inferences.

Using the terminology of *the Standards* (1999/2014), the interpretive argument used sources of evidence based on test content, including (a) expert-informed content domain description, (b) logical analyses of the correspondence between test content and test domain, including creating test specifications, and (c) expert reviews for test content and test task relevance. Adequate reliability indices were a source of validity evidence based on test internal structure. Next, evidence about relations to other variables was based on experimental analyses of discriminant relationships with a different construct (i.e., relationship between a visual-inclusive construct and an audio-centric construct). The final source of evidence was stakeholders' perceptions of the visually content-rich assessment construct. Although there is yet no clearly designated place for this source in the Standards, it seems it could fit under evidence based on relations to other variables, with stakeholders' perceptions serving as one of the "other variables."

This study does not claim to have developed a fully-fledged validity argument for the video-based ALC test. A complete validity argument would require assessing extrapolation and utilization inferences as well as gathering more evidence for the explanation inference. However, it has laid the groundwork for validating L2 academic listening tests that contain academically relevant videos. It collected evidence to argue that L2 academic listening tests are more valid with content-rich videos than without. Although this argument developed based primarily on evidence for the explanation inference, backing for the other above-mentioned inferences bears significance.

Table 5.1

Interpretive Argument for the Video-Based ALC Test

Inference	Warrant	Assumptions	Backings	Literature, logical evidence, and Statistical analyses
Domain definition	Test-takers' performance reflects their skills in authentic academic listening contexts.	<ul style="list-style-type: none"> a) Language skills and processes needed for the TLU are identified. b) Assessment tasks typical of the TLU are identified. c) Test tasks can be created to reflect (a) and (b) 	<ul style="list-style-type: none"> a) Theory-driven identification of the required language abilities b) Expert-based identification of typical TLU assessment tasks c) Expert-driven development of test content and tasks 	<ul style="list-style-type: none"> a) Field (2013) b) Field (2009; 2011); Powers (1985) Richards (1983) c) Appropriateness of test content and tasks was judged by experts in L2 teaching and assessment; multiple rounds of review and revision
Evaluation	Observed scores reflective of the academic listening ability are consistently awarded.	<ul style="list-style-type: none"> a) Appropriate scoring rubrics or methods are used. b) Test performance is not affected by administration conditions. c) Psychometric properties of items are appropriate for norm-referenced decisions. 	<ul style="list-style-type: none"> a) Clear rubrics and reliable scoring methods b) Testing conditions are properly controlled c) Item analysis and descriptive statistics 	<ul style="list-style-type: none"> a) Automated scoring method; answer key checked for correctness b) Set of standardized instructions before and during test-taking; elimination of contaminated/problematic responses c) Rasch item reliability analysis; Rasch fit statistics; classical analyses of item difficulty and discrimination
Generalization	Test scores are generalizable to expected scores in authentic academic listening contexts.	<ul style="list-style-type: none"> a) The number of items is sufficient for stable estimates of test-takers' performance. b) Test administration can be easily replicated for other samples. 	<ul style="list-style-type: none"> a) Reliability and generalizability studies b) Evidence based on test specifications 	<ul style="list-style-type: none"> a) Rasch person reliability index and Cronbach's alpha indices by mode b) Detailed test specification and item writing techniques
Explanation	Observed scores are attributed to the academic listening construct.	<ul style="list-style-type: none"> a) Item difficulty is affected by construct-relevant factors (e.g., item type). b) The construct definition is supported by test stakeholders. 	<ul style="list-style-type: none"> a) Analysis of the effect of content-rich visuals on item difficulty b) Analysis of stakeholders' perceptions about the visually content-rich academic listening construct 	<ul style="list-style-type: none"> a) Rasch and classical analyses on item difficulty differences (individually or collectively) by delivery mode, accounting for proficiency level, video-dependence, and listening subskill; correlational analyses of testlet difficulty and self-reported viewing behavior b) Test-takers' and teachers' perceptions of a visually content-rich academic listening construct

Implications

This study has several implications for the field of L2 assessment. This section describes theoretical and methodological implications first, followed by practical recommendations for L2 academic listening assessment and pedagogy.

Theoretical implications. Results of this study contributed to existing theories on academic listening and test validation. First, the study provided strong support that the construct of L2 academic listening is inclusive of content-related visual information. The study concluded that content-rich visual information was part of academic listening contexts (Lynch, 2011; Field, 2009; 2011). In addition, findings empirically showed that content-rich visuals were construct-relevant. Positive effects of content-rich videos observed in the present study were theoretically grounded. These findings are in agreement with the latest tendency to conceptualize the L2 listening skill as a process integrating linguistic and outside knowledge, kinesic information, and exophoric visuals (e.g., Field, 2008; Flowerdew & Miller, 2005; Rost, 2016; Vandergrift & Goh, 2012; Ur, 1984). Understanding visual nuances of the construct is vital because it leads to more refined definitions, more valid assessments, and more effective pedagogies of L2 academic listening comprehension.

Next, this study offers an example of applying the argument-based validity framework to video-based listening tests. The essence of this framework is backing the six validity inferences (i.e., test domain, evaluation, generalization, explanation, extrapolation, and utilization) with empirical evidence. This study was grounded in test domain while also considering evaluation and generalization inferences; it focused on gathering evidence for the explanation inference, with the intent to justify the inclusion of

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

content-rich videos into the L2 academic listening construct. Traditionally, backing for the explanation inference came from studies employing correlational analysis, factor analysis, group comparison, and cognitive processing analysis to detect and eliminate construct-irrelevant variance (Chapelle et al., 2008; Kane, 2004; 2006; 2013). The present study complemented these methods. Following Gruba (2014) and Bachman & Palmer (2010), it highlighted the importance of evidence stemming from test stakeholders' perceptions of the construct for backing the explanation inference. Test-takers' and teachers' perceptions were elicited to gauge the effects of content-rich videos on listening comprehension difficulty, motivation, and authenticity. Test-takers and teachers expressed their agreement with including content-rich videos in L2 academic listening tests. As long as hypotheses for stakeholders' perceptions are theory-informed, construct-relevant, and include additional sources of test validity evidence from *the Standards* (1999/2014), these methods enrich the existing validity framework by strengthening the explanation inference.

Methodological implications. This study offers three methodological implications concerning the choice of statistical analyses, visual classification, and item classification, each of which is described below.

Recall that this study employed both classical and statistical analyses for detecting the effects of content-rich videos on L2 academic listening comprehension. Seven classical analyses of variance (ANOVA) were used to compare the video-based and audio-only modes on video-dependent items (local and global) and video-independent items (local and global), and overall. Mode and proficiency were the two factors in each

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

ANOVA. Then, Rasch analyses were run on the same groups of items and overall, including one facet measurement report and five interactions.

The use of both the classical and Rasch methods allowed for comparing the insights they provided into the effects of content-rich videos versus just audio on L2 academic listening comprehension. While the Rasch approach is commonly employed for item analysis, its use for group comparison studies may still be limited. It was believed that comparing the two approaches would show whether Rasch analysis is comparable and perhaps preferable to classical methods in terms of estimating group differences, which is discussed below.

Many-faceted Rasch measurement (MFRM) analysis may offer a better alternative for detecting effects of mode and proficiency on items than factorial ANOVA analyses. In this study, video-independent items were harder with videos than without, as showed by MFRM interactions. More specifically, video-independent global items were harder in the video-based mode than in the audio-only mode for lower-level students. Conventional ANOVAs did not find these effects. This may reflect the popular view that MFRM is more informative than analyses based on the classical test theory (e.g., Batty, 2015; Wright, 1992). An often-cited advantage of MFRM is its capacity to estimate person abilities and item difficulties based on truly continuous scales, which leads to “more principled comparisons” (Batty, 2015, p. 9). Also, unlike raw scores in the conventional analyses, MFRM takes difficulty of items into account when estimating persons’ abilities. For example, a test-taker getting hard items correct would have a higher ability estimate compared to a test-taker who succeeds on many easy items but misses hard items. The same principle is applied to MFRM item difficulty estimation: An

item passed by many test-takers may still be estimated as difficult if these test-takers had high abilities (Bond & Fox, 2015; McNamara, 1996). Perhaps due to these advantages, the MFRM in the present study was more informative than the classical raw-score-based ANOVA analyses.

One may argue that more sophisticated classical analyses could provide insights and precision similar to Rasch estimators. This study used fixed-effects factorial ANOVAs, which are relatively basic statistical procedures. Still, the results of the factorial ANOVAs and Rasch interactions in this study overlapped, providing similar findings on many groups of items. This may suggest that more advanced statistical models, such as random or mixed-effects factor models with proficiency as a covariate, could be sufficiently effective in discovering differences due to delivery mode and listening proficiency on different groups of items. Future studies can confirm this supposition by comparing the effectiveness of various statistical techniques based on classical test theory against Rasch analysis.

Some investigators may prefer classical over Rasch analyses because of the interpretability issues. Classical analyses normally provide results on a scale that is relative to the instrument's total score. This allows not only for interpreting comparisons of interest but also for estimating the magnitude of test-takers' success in a particular condition with ease. For example, test-takers' ALC test average scores of 11.53 and 13.10 in the video-based and the audio-only conditions respectively seem more informative than the corresponding Rasch estimators of 0.07 and -0.07 (see Tables 4.10 and 4.24 in Chapter 4). The classical scores immediately connote with the highest possible score, thereby informing the reader of how easy the test was, which makes the

overall comparison more meaningful. While the Rasch estimators can also be related to the highest person ability logit in the dataset, this link may not be as easily interpretable. Therefore, classical analyses may do a better job communicating results to the reader.

Despite its lower interpretability, MFRM may still be preferred to classical test analyses of group comparisons on different sets of items. This study investigated seven sets of ALC items, namely all 24 items, 14 video-dependent items (7 local and 7 global), and 10 video-independent items (5 local and 5 global). To estimate the effects of delivery mode and listening proficiency, seven separate ANOVAs had to be run because there was no way to account for performance on different sets of items within one omnibus analysis. In contrast, Rasch analysis could handle investigations into the effects on different item sets within one specified command file. It also provided estimations of the effects on each individual item, which is still a challenge for traditional classical analyses, such as ANOVA or regression. Considering this and other advantages of Rasch analysis, it is recommended for use in group comparison contexts as a supplement to classical analyses, if not as an alternative.

Finally, Rasch analysis offers additional potential for estimating psychometric properties of test items. It can provide test developers with infit values, showing how well each item fits into the construct. In this study, infit values for each item were within the norm while classical item discrimination indices for some items were beyond the expected range (see Table 4.5 on p. 141). This may show that item infit estimation is a valuable addition to psychometric analyses, as it reveals information about items that is not reflected by classical estimators. In addition, a Rasch person-ability, or Wright map, is a very useful psychometric tool. Placing person abilities and item difficulties on the

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

same scale gives a concise but sufficient picture of how well the test suits test-takers, which is not always afforded by classical methods. According to classical item difficulties in this study, items ranged in difficulty from 0.30 to 0.83, suggesting that the test suited test-takers of ranging proficiency levels (see Table 4.5 on p. 141). However the Wright map on p. 144 (see Figure 4.2) revealed the lack of higher-difficulty items. To draw a fuller picture of item functioning for a multiple-choice test, it is, therefore, strongly recommended that classical item difficulty and item discrimination analyses should be complemented with Rasch infit and person ability estimators.

The second methodological implication of this study concerns classifications of videos for L2 academic listening tests. Previous research used the distinction between context and content visuals for examining the effects of video type on comprehension (Bejar et al., 2000; Ginther, 2002; Suvorov, 2015a; 2015b). This study discussed the shortcomings of this classification and proposed a new video type, coined as content-rich videos (see pp. 29-32 in Literature Review). Content-rich videos were thoroughly defined. The definition specified patterns of content-related visuals in a video in terms of amount, kind, functions, congruity with the auditory stimulus, and interpretability. This newly-defined video type reflected visual patterns found in the selected authentic lecture passages, which worked properly for the purposes of this study. Although it cannot be taken as a universal visual representation of authentic academic contexts, the content-rich video type illustrates a new approach that seems to be more stable and more informative for future research, if not yet for existing large-scale standardized tests. It can also inspire more elaborate and more effective future classifications of videos.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

The third methodological implication is a new classification for multiple-choice items testing academic listening comprehension and beyond. The classification distinguishes between video-dependent and video-independent test items. Video-dependent items are cued by video-based visuals while video-independent items are not. Determination of video-dependence for an item involved several rounds of piloting, administering a video-dependence questionnaire, and administering the muted version of the test. Thus, this study provides a framework for classifying individual test items based on their relationship to video-based visual information as well as potential techniques to apply this classification. This framework may guide future studies about visual effects on comprehension and be the gateway for test developers to develop more accurate specifications for video-based tests or testlets.

Assessment implications. From the assessment standpoint, the study offers three recommendations. First and foremost, test developers are urged to use content-rich videos in their L2 academic listening tests. The absence of visual information typical of academic contexts was shown to diminish test authenticity and unfairly increase test difficulty. In addition, audio-only tests may not be as motivating for test-takers as video-based tests, as suggested by L2 teachers in this study. Authenticity, difficulty, and motivation are all construct-relevant factors, along with content-rich visuals themselves (e.g., Chapelle et. al., 2008; Flowerdew & Miller, 2005; Li, 2013; Rost, 2016; Tafaghodtari & Vandergrift, 2008; Vandergrift, 2005). Viewing visuals is a process that also constitutes a construct-relevant factor (Wagner, 2007; 2008; 2010a). In this dissertation study, viewing behavior was part of the listening process and related to test-takers' listening comprehension. Using visual-free L2 academic listening tests would

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

remove all these variables from the equation, thereby undermining the measured construct. To avoid this, it is suggested that L2 assessment companies start using content-rich videos in L2 academic listening tests. In the age of new media, it is a progressively lesser challenge than before and should no longer get in the way of developing more valid tests (Gruba, 2014).

This dissertation may serve as one means of encouraging test developers to include content-rich videos in their L2 academic listening tests. Another means of doing so would be to publish the findings of this study in influential professional journals in the field of L2 assessment, such as *Language Testing* and *Language Assessment Quarterly*. Similarly, research reports based on this study can be published online on the websites of testing companies that funded this research, namely *The Paragon Testing Enterprises* (Canada), *the British Council* (UK), and *the Educational Testing Service* (USA). Other venues for disseminating this study's findings are giving presentations at professional assessment-oriented conferences, such as *the Language Testing Research Colloquium*, *the Language Assessment Research Conference*, *the East Coast Association of Language Testers*, and *the European Association for Language Testing and Assessment*. Lastly, intensive English language programs (IEP) in the USA can be approached as a launching ground for video-based academic English testlets. The IEPs at some universities targeting academic English (e.g., the Program in Intensive English at Northern Arizona University) might be interested in including one or more testlets used in this study into their placement tests. This practice could be further extended to foreign IEPs, thereby increasing the impact of this study on the field of academic L2 listening assessment.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Second, the inclusion of content-rich videos in listening tests necessitates the use of video-dependent items. In authentic contexts, students use content-related visual information to understand detail-oriented or overall lecture points (Lynch, 2011). Using video (along with audio) for answering individual local and global test items would be a reflection of this. While developing video-dependent items entails extra work and research, it was shown to be a relatively straightforward and replicable process. It may include training for the development of video-cued items and piloting these items with test stakeholders, with the intention to confirm the items' video-dependent design. Results of this study suggested having multiple rounds of such piloting, as some video-dependent items functioned unexpectedly in the video-based mode despite having been initially piloted. To ensure that comprehension on individual items is not misguided by videos, larger-scale product-oriented and some process-oriented data from test-takers would be helpful. Evidence from several sources would best inform the process of developing video-dependent items.

Third, video-based L2 academic listening testlets could exclude video-independent items for the time being. Content-rich videos may introduce bias against lower-level students on global video-uncued items, as indicated by results of this study. At this point, it is unclear if this bias was caused by content-related visuals (i.e., graphs and pictures) or by the lecturer's non-verbal cues, or by both. Considering Ginther's (2002) recommendation to use visuals as long as they do not hurt comprehension, one could argue for refraining from using traditional video-independent items for video-based academic listening passages until further research is done.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

This need not entail using exclusively video-cued items in an entire test, however, as this could inflate the role of videos in comprehension. Even authentic contexts may occasionally rely on mostly the auditory information (e.g., a PowerPoint-free lecture introduction), which suggests that a test should also include some items uncued by content-related visuals. Alternating between video-based and audio-only testlets within a test may be a workaround to this dilemma, with content-rich videos going with video-dependent items and audio-only testlets using traditional items. Having a mix of audio-only and video-based testlets would allow for improving validity of the test while avoiding bias against lower-level test-takers.

On the other hand, this “bias” against lower-level test-takers might be natural. The common wisdom suggests that some comprehension questions in many authentic lectures rely on visual information while others do not. Using the terminology of this study, authentic university lectures may use both video-dependent and video-independent comprehension questions. From this perspective, the inclusion of video-independent items along with video-dependent items in video-based testlets may be supportive of the test domain inference in the interpretive argument. Moreover, video-dependent and video-independent items may be said to measure the same construct. Recall that the construct of L2 academic listening comprehension was defined as “the active process of receiving and constructing meaning from the spoken lecture input, the lecturer’s non-verbal cues, situational cues, and content-rich visual aids with the help of note-taking” (see p. 50 in Literature Review). Video-dependent items relied mostly on the “content-rich visual aids” part of this definition. Video-independent items relied mostly on “the spoken lecture input.” Both types were, to some extent, related to the non-verbal cues and

situational cues in the video, although not to the point where non-verbal and situational visuals would provide or strongly allude to the correct answer. We can see that both types of items were affiliated with the same construct, which also may provide support for including both item types within video-based L2 academic listening testlets. Therefore, it is recommended that video-independent items be used alongside video-dependent items in video-based listening testlets until counterevidence is produced.

Pedagogical implications. The study has two implications for the field of L2 academic listening pedagogy. First, the popular tendency to use authentic video-based materials in L2 listening classrooms was supported as a way to activate schemata, exercise listening processes, improve listening motivation, and bring in the sociolinguistic dimension of listening (Field, 2013; Flowerdew & Miller, 2005; Vandergrift, 2004). This study further suggested that academic listening classrooms should incorporate videos in their assessments and self-assessments. For example, self-evaluating comprehension-checking activities based on authentic lecture videos would activate the process of listening while also evaluating students' academic listening ability more accurately (Chapelle & Jamieson, 2008; Flowerdew & Miller, 2005).

According to Chapelle and Jamieson (2008), using computerized assessments with the possibility to get immediate feedback on test items could "heighten students' awareness of their understanding" (p. 142). This prospect becomes more valuable in the case of tests with content-rich videos. In addition to increasing listening self-awareness, it could lead test-takers to re-evaluate the potential of visual information to aid listening comprehension. To resolve an automatically identified error, a student would be more likely to replay relevant parts of the video and find the correct answer with the help of

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

video, assuming the student has access to video controls. In other words, instant feedback in low-stakes computerized video-based tests can raise awareness of both the listening process itself and the facilitative role of content-related visuals in this process. Using such assessments in classroom practice activities is recommended.

Second, classroom activities should heighten students' awareness of the roles of viewing behavior and visual literacy in academic listening. This study and others showed that viewing behavior is part of academic listening comprehension (Wagner, 2007; 2010a; Suvorov, 2013; 2015a). Similarly, visual literacy was shown to be an important factor for successful comprehension (e.g., Beaudoin, 2016; Malamitsa, Kokkotas, & Kasoutas, 2008). Test-takers who are not skilled in viewing behavior and have low visual literacy may have difficulties succeeding in academic listening comprehension. Thus, it may be worthwhile to integrate activities for developing viewing behavior skills and visual literacy in L2 academic listening classrooms.

While modern English for academic purposes (EAP) teaching materials build on multimodal resources, teaching methodologies for EAP listening or L2 academic listening in general give little focus to visual literacy (Chun, 2015). Activities specifically targeting visual skills yet tailored for L2 contexts should be developed, researched, and put into practice. Such activities would focus on developing test-takers' abilities to identify visual information and interpret its meaning, analyze and evaluate multimedia texts, and use visual conventions, among others (e.g., Avgerinou & Pettersson, 2007; Hattwig, Bussert, Medaille, & Burgess, 2012; "NCTE Framework," 2013; "ACRL Standards," 2011). It would also be advisable to consult existing literature on multimodal teaching and learning methods in core curriculum teaching contexts (e.g., Kress, 2010;

Kress et al., 2005). They may guide L2 teaching concerning visual displays, spatial arrangements, time management, and other factors that can foster visual learning in pedagogical environments. Integrating visuals-focused activities in L2 classrooms should help students become more skillful academic listeners, better prepared for university life in English-speaking countries.

Limitations of the Study

This study was limited in four ways. First, there was no proctoring. The academic listening comprehension test, the anchor test, and the test-takers' questionnaire were administered online. While test-takers were urged to avoid distractions and remain attentive, their proper testing behavior was not guaranteed. Test-takers were not directly observed while taking the assessment instruments, which may have caused undesired variance in test-takers' responses.

Second, the authenticity segment of the test-takers' questionnaire was flawed. It asked: How realistic was this lecture? It was expected that test-takers in the audio-only mode would give lower ratings for this item than test-takers in the video-based mode on average. However, the audio-only group may have still considered audio-only lectures highly authentic simply because they had not been exposed to video-based tests before and, thus, had no frame of reference in relation to which their judgement could be made.

Third, the sample of test-takers was not balanced. There was a high proportion of Spanish-speaking test-takers, which may not accurately reflect the population of ESL and EFL learners worldwide. Therefore, the results of this study cannot be generalizable to this population, but should rather be viewed as a valuable source of information about ESL and EFL learners worldwide.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Finally, the anchor test did not have the capacity to distinguish between more than two levels of academic listening proficiency. It precluded this study from examining the role of content-rich videos for low beginner and high advanced learners' comprehension. It was only possible to differentiate between two broader proficiency categories, namely lower and higher.

Directions for Future Research

Several suggestions for future investigations into visually content-rich video-based assessments can be made. This section presents these suggestions in relation to corresponding validity inferences.

Regarding the test domain inference, more work should be done to develop the definition of content-rich videos. This dissertation study selected four authentic YouTube lecture videos based on pre-specified criteria. Then, it imitated visual patterns found in these videos while adjusting them so that they would be similar across lectures. While this method helped to contrive videos reflective of the original lectures, the extent to which the original lectures themselves were typical of academic contexts visual-wise was unknown. Therefore, future studies may need to conduct a comprehensive review or analysis of video lectures available on YouTube and other online platforms to arrive at the conclusion as to what a typical visual pattern is with respect to types, configurations, and amount of visual information in a lecture. This may also include analyses of rhetorical effectiveness and content dynamism, as suggested by Suvorov (2013).

The explanation inference may be strengthened by including analysis of video effects and stakeholders' perceptions on academic listening comprehension by testlet. This study combined the scores on each testlet, treating each testlet as a blocking factor.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

While contributing to variability of test and questionnaire scores, testlets as sources of variation were not of interest to this study. Investigating individual testlets could provide additional insights into how graphs and pictures affect L2 academic listening comprehension. It would allow for including additional variables, such as register (e.g., hard vs. soft science lectures), length, speech rate, and topic.

The interpretive argument in this study clearly lacked process-oriented evidence. Analysis of test-takers' response processes could uncover individual variance that is often masked when focusing on group performance (Wagner, 2013, p. 180). It could also provide support for the assumption that test-takers utilize visual decoding processes and higher-level processes described in Field's model (2013) for lecture comprehension. Such analyses would offer insights as to how test-takers use those processes for answering comprehension questions, possibly supporting the quantitative results of decreased comprehension difficulty on video-dependent items and providing explanations for the unexpected results of increased comprehension difficulty on video-independent items in the video-based mode. This can be done using verbal reports, verbal protocols, or interviews (Bachman & Palmer, 2010; Green, 1998). Process-related evidence would serve as a strong backing for the explanation inference in the validity argument for including content-rich videos in L2 academic listening tests.

Next, the interpretive argument in this study did not examine the ALC test's relationships to other measures of a similar construct. Such an investigation would necessitate finding an established academic listening test with content-related videos and correlating the performances on both tests. High correlation coefficients would provide further evidence for the explanation inference in the argument. This scenario is somewhat

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

problematic because there are currently few or no reputable academic English listening tests on the market that would operationalize L2 academic listening as a visually inclusive skill (Kang, Gutierrez Arvizu, Chaipupae, & Lesnov, 2016). Researchers are left with opportunities to correlate video-based tests with existing audio-centric measures of listening proficiency, expecting to find moderate associations. For example, test-takers' scores on a video-based academic listening test could be correlated with test-takers' scores on existing video-free high-stakes tests, such as TOEFL or IELTS. A moderate correlation coefficient would function as an additional backing for the explanation inference.

The present study leaves the extrapolation and the utilization inferences unassessed. An analysis of predictive relationships between the ALC test and a criterion in the TLU domain, such as academic achievement, would be one way to further assess the extrapolation inference. If the ALC test correlates higher with academic achievement than an audio-only test does, it will signal a stronger potential of the video-based ALC to predict test-takers' behavior in the TLU domain. This would strengthen the overall validity argument in favor of using content-related visuals in the listening tests, providing evidence for one more building block in the chain of validity inferences.

Evidence for intended and unintended consequences of using the ALC test was also missing. It would include investigating how useful test scores are for making decisions about test-takers' academic listening ability (e.g., test scores are easily interpretable) and whether the test has a positive impact on how L2 is taught (e.g., visual literacy becomes a student learning outcome for L2 academic listening classes). Future consequence-oriented investigations of using content-rich videos in L2 academic

listening tests could provide evidence for the tests' proper interpretations and positive washback effects on teaching and learning, thereby supporting the utilization inference.

Conclusion

This dissertation study paved the way for including content-rich videos in L2 academic listening tests. It showed that tests with content-rich videos would be more representative of authentic contexts. Empirical evidence demonstrated that the video-based mode affected test-takers' performance in accord with theoretical expectations. Test-takers' and teachers' perceptions supported the visual-inclusive construct and challenged the audio-centric construct. Finally, all this evidence for the domain definition, evaluation, generalization, and explanation inferences in the interpretive argument supported the use of content-rich videos in tests.

The argument presented in this study is yet to be completed. However, it is substantive enough to attract the attention of high-stakes test developers and assessment researchers. It is hoped that future investigations will adopt and expand on the ideas from the present study, furthering the understanding of the L2 academic listening assessment construct. It is hoped that evidence from these future studies and the present dissertation study will move test developers to reconsider the benefits of including content-rich videos in tests and initiate the long-awaited practice of making more valid L2 academic listening tests.

References

- ACRL Visual Literacy Competency Standards for Higher Education. (2011, October). Retrieved from <http://www.ala.org/acrl/standards/visualliteracy>
- Aiken, E., Thomas, G., & Shennum, W. (1975). Memory for a lecture: effects of notes, lecture rate, and informational density. *Journal of Educational Psychology*, 67, 439-444.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- An ILA definition of listening (1995). *ILA Listening Post*, 53, 1.
- Anderson, J. (2000). *Cognitive psychology and its implications* (5th ed.). New York, NY: Worth Publishers.
- Anderson, N. (2014). *Empirical direction in design and analysis*. New York, NY: Routledge.
- Aryadoust, V. (2013). *Building a validity argument for a listening test of academic proficiency*. Cambridge, UK: Cambridge Scholars Publishing.
- ASTR105x - Alien Worlds (2014, November 24). *ASTR105x_M06_Lecture* [Video File]. Available from <https://www.youtube.com/watch?v=cYmBoAEQtQI>
- Avgerinou, M., & Pettersson, R. (2011). Toward a cohesive theory of visual literacy. *Journal of Visual Literacy*, 30, 1-19.
- Bachman, L. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2, 1-34.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

- Bachman, L., & Palmer, A. (1996). *Language testing in practice: Developing language assessments and justifying their use in the real world*. Oxford, UK: Oxford University Press.
- Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford, UK: Oxford University Press.
- Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transactions*, 22, 1145-1146.
- Baltova, I. (1994). The impact of video on the comprehension skills of core French students. *Canadian Modern Language Review*, 50, 507–531.
- Barlet, F. (1932). *Remembering*. Cambridge, UK: Cambridge University Press.
- Batty, A. (2015). A comparison of video- and audio-mediated listening tests with many-facet Rasch modeling and differential distractor functioning. *Language Testing*, 32, 3-20.
- Beaudoin, J. (2016). Describing images: A case study of visual literacy among library and information science students. *College & Research Libraries*, 77, 376-392.
- Becker, A. (2016). L2 students' performance on listening comprehension items targeting local and global information. *Journal of English for Academic Purposes*, 24, 1-13.
- Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper* (TOEFL Monograph Series Report No. 19). Princeton, NJ: Educational Testing Service.
- Benson, M. (1994). Lecture listening in an ethnographic perspective. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 181-198). New York, NY: Cambridge University Press.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Blanca, M., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-normal data:

Is ANOVA still a valid option? *Psicothema*, *29*, 552-557.

Bodie, G., Janusik, L., & Valikoski, T.-R. (2008). Priorities of listening research: Four interrelated initiatives. *A white paper sponsored by the Research Committee of the International Listening Association*. Retrieved from

http://www.listen.org/resources/documents/white_paper_prioritiesresearch.pdf

Bond, T., & Fox, C. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge.

Brett, P. (1997). A comparative study of the effects of the use of multimedia on listening comprehension. *System*, *25*, 39-53.

Brindley, G. (1998). Assessing listening abilities. *Annual Review of Applied Linguistics*, *18*, 171-191.

Brown, J. D. (2000, October). What is construct validity? *JALT Testing & Evaluation SIG Newsletter*, *4*, 8-12.

Brunye, T., Taylor, H., & Rapp, D. (2008). Repetition and dual coding in procedural multimedia presentations. *Applied Cognitive Psychology*, *22*, 877-895.

Buck, G. (2001). *Assessing listening*. Cambridge, UK: Cambridge University Press.

Carrell, P., Dunkel, P., & Mollaun, P. (2002). *The effects of notetaking, lecture length and topic on the listening component of the TOEFL 2000* (TOEFL Monograph Series No. MS-23). Princeton, NJ: ETS.

Chafe, W. (1979). The flow of thought and the flow of language. In T. Givon (Ed.), *Syntax and semantics, 12: Discourse and syntax* (pp. 159-181). New York, NY: Academic Press.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

- Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks, and test construction. *Language Testing*, 14, 3-22.
- Chalhoub-Deville, M. (2003). Second language interaction: current perspectives and future trends. *Language Testing*, 20, 369-383.
- Chapelle, C. & Jamieson, J. (2008). *Tips for teaching with CALL: Practical approaches to computer-assisted language learning*. White Plains, NY: Pearson Education.
- Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. Bachman & A. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32-70). Cambridge, UK: Cambridge University Press.
- Chapelle, C. (2011). Validation in language assessment. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 717-730). New York, NY: Routledge.
- Chapelle, C. (2012). Validity argument for language assessment: The framework is simple... *Language Testing*, 29, 19-27.
- Chapelle, C., & Voss, E. (2014). Evaluation of language tests through validation research. In J. Kunnan (Ed.), *The Companion to language assessment* (pp. 1081-1097). London, UK: John Wiley.
- Chapelle, C., Enright, M., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.
- Chapelle, C., Enright, M., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, 29, 3-13.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Chappelle, C. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, 19, 254-272.

Chastain, K. (1976). *Developing second-language skills: Theory to practice*. Chicago, IL: Rand McNally College Pub. Co.

CheckMarket. Sample Size Calculator [Computer Software]. (2017). Available from <https://www.checkmarket.com/sample-size-calculator/>

Chun, C. (2015). *Power and meaning making in an EAP classroom. Engaging with the everyday. Critical language and literacy studies: 19*. Bristol, UK: Multilingual Matters.

Clark, H. & Clark, E. (1977). *Psychology and language*. New York, NY: Harcourt Brace Jovanovich.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Collis, B., & Wende, M. (2002). *Models of technology and change in higher education. An international comparative survey on the current and future use of ICT in higher education* (Report). Retrieved from <https://pdfs.semanticscholar.org/b8a8/afc75551c3e8eaeaf1fbeb52174d05c7d3c8.pdf>

Coniam, D. (2001). The use of audio and video comprehension as an assessment instrument in the certification of English language teachers: A case study. *System*, 29, 1-14

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

- Cook, J. (1975). *A communicative approach to the analysis of extended monologue discourse and its relevance to the development of teaching materials for ESP*. (Unpublished master's thesis). University of Edinburgh, UK.
- Copyright Law of the United States and Related Laws Contained in title 17 of the United States Code. (2011). Retrieved from <http://www.copyright.gov/title17/circ92.pdf>
- Creswell, J. (2013). *Qualitative inquiry and research design: Choosing among five approaches* (3rd ed.). Thousand Oaks, CA: Sage Publications.
- Cronbach, L. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.
- Cubilo, J., & Winke, P. (2013). Redefining the L2 listening construct within an integrated writing task: Considering the impacts of visual-cue interpretation and note-taking. *Language Assessment Quarterly*, 10, 371–397.
- Cumming, J., & Maxwell, G. (1999). Contextualising authentic assessment. *Assessment in Education*, 6, 177–94.
- Cutler, A., & Clifton, Jr., C. (1999). Comprehending spoken language: A blueprint of the listener. In C. Brown, & P. Hagoort (Eds.), *The neurocognition of language* (pp. 123-166). Oxford, UK: Oxford University Press.
- Daniel, J. (2011). *Sampling Essentials: Practical guidelines for making sampling choices*. Los Angeles, CA: Sage Publications.
- Debes, J. L. (1968). Some foundations for visual literacy. *Audiovisual Instruction*, 13, 25-27.
- DeCarrico, J., & Nattinger, J. (1988). Lexical phrases for the comprehension of academic lectures. *English for Specific Purposes*, 7, 91-102.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

- Dewey, J. (1929). *The quest for certainty*. New York, NY: Minton Balch And Company.
- Dorans, N., Moses, T., & Eignor, D. (2010). *Principles and practices of test score equating* (TOEFL Research Rep. No. ETS RR-10-29). Princeton, NJ: Educational Testing Service.
- Dornyei, Z. & Taguchi, T. (2009). *Questionnaires in second language research*. New York, NY: Routledge.
- Douglas, D. (1997). Language for specific purposes testing. In C. Clapham, & D. Carson (Eds.), *Encyclopedia of language in education. Volume 7: Language testing and assessment* (pp. 111-120). Dordrecht, Netherlands: Kluwer Academic.
- Dudley-Evans, A., & Johns, T. (1981). A team teaching approach to lecture comprehension for overseas students. In T. Dudley-Evan, & T. F. Johns (Eds.), *The Teaching of listening comprehension* (pp. 30-6) (ELT Documents Special). London, UK: The British Council.
- Dudley-Evans, T. (1994). Variations in the discourse patterns favoured by different disciplines and their pedagogical implications. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 146-158). New York, NY: Cambridge University Press.
- Dunkel, P. (1988). The content of L1 and L2 students' lecture notes and their relation to test performance. *TESOL Quarterly*, 22, 259-282.
- ETS Guidelines for Fair Test and Communication*. (2015). Princeton, NJ: Educational Testing Service.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160.
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191.
- Field, J. (2008). *Listening in the language classroom*. Cambridge, UK: Cambridge University Press.
- Field, J. (2009). A cognitive validation of the lecture-listening component of the IELTS listening paper. In L. Taylor (Ed.), *IELTS research reports* (Vol. 9) (pp. 17-66). Canberra: IELTS Australia, Pty Ltd & the British Council.
- Field, J. (2011). Into the mind of the academic listener. *Journal of English for Academic Purposes, 10*, 102-112.
- Field, J. (2013). Cognitive validity. In A. Geranpayeh, & L. Taylor (Eds.), *Examining listening. Research and practice in assessing second language listening* (pp. 77-151). Cambridge, UK: Cambridge University Press.
- Fink, A. (2009). *How to conduct surveys* (4th edition). Thousand Oaks, CA: SAGE Publications.
- Fisher, W. (2008). Cash value of reliability. *Rasch Measurement Interactions, 22*, 1160.
- Flowerdew, J. (1994). Research of relevance to second language lecture comprehension: An overview. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 7-29). New York, NY: Cambridge University Press

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

- Flowerdew, J., & Miller, L. (2005). *Second language listening: Theory and practice*. New York, NY: Cambridge University Press.
- Fowler, F. (2014). *Survey research methods*. Thousand Oaks, CA: SAGE Publications.
- Fries, C. (1947). *Teaching and learning English as a foreign language*. Ann Arbor, MI: University of Michigan Press.
- Fulcher, G. (2010). *Practical language testing*. London, UK: Hodder Education.
- Furness, E. (1952). Techniques for the teaching of listening. *The Modern Language Journal*, 36, 124-128.
- G*Power 3.1.9.3 [Computer Software]. (2014). Available from http://www.gpower.hhu.de/fileadmin/redaktion/Fakultaeten/Mathematisch-Naturwissenschaftliche_Fakultaet/Psychologie/AAP/gpower/GPowerMac_3.1.9.3.zip
- Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Language Testing*, 19, 133–167.
- Glass, G., Peckham, P., & Sanders, J. (1972). Consequences of failure to meet assumptions underlying fixed effects analyses of variance and covariance. *Review of Educational Research*, 42, 237-288.
- Green K. (1998). The use of auditory and visual information during phonetic processing: Implications for theories of speech perception. In R. Campbell & B. Dodd (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech* (pp. 3-25). London, UK: Erlbaum.
- Green, A. (1998). *Studies in Language Testing 5: Verbal protocol analysis in language testing research*. Cambridge, UK: Cambridge University Press.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

- Gruba, P. (1993). A comparison study of audio and video in language testing. *JALT Journal*, 15, 85–88.
- Gruba, P. (2004). Understanding digitized second language videotext. *Computer Assisted Language Learning*, 17, 51-82.
- Gruba, P. (2014). New media in language assessment. In J. Kunnan (Ed.), *The Companion to language assessment* (pp. 995-1012). London, UK: John Wiley.
- Ha, R., & Ha, J. (2012). *Integrative statistics for the social and behavioral sciences*. Thousand Oaks, CA: Sage.
- Haladyna, T. M., Downing S. M., Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15, 309-334.
- Hansen, C., & Jensen, C. (1994). Evaluating lecture comprehension. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 241-268). New York, NY: Cambridge University Press.
- Hatch, E. & Lazaraton, A. (1990). *The research manual: Design and statistics for applied linguistics*. Boston, MA: Heinle & Heinle Publishers.
- Hattwig, D., Bussert, K., Medaille, A., & Burgess, J. (2012). Visual literacy standards in higher education: New opportunities for libraries and student learning. *Libraries and the Academy*, 13, 61-89.
- Hilpert, M. (2014, January 13). *First language acquisition*. [Video File]. Retrieved from <https://www.youtube.com/watch?v=up0yVJWf9zQ>
- Hulstijn, J., Young, R., & Ortega, L. (2014). Bridging the gap: Cognitive and social approaches to research in second language learning and teaching. Editor's

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

- introduction & Editor's closing thoughts. *Studies in Second Language Acquisition*, 36, 361-365.
- IBM Corp [Computer Software]. (2016). IBM SPSS Statistics for Windows, Version 24.0. Armonk, NY: IBM Corp.
- iMovie (Version 10.1.1) [Computer Software]. (2017). Available from <https://www.apple.com/imovie/>
- Jamieson, J. (2014). Defining constructs and assessment design. In J. Kunnan (Ed.), *The Companion to language assessment* (pp. 769-787). London, UK: John Wiley.
- Jordan, R. (1997). *English for academic purposes. A guide and resource book for teachers*. Cambridge, UK: Cambridge University Press.
- Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527-535.
- Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research & Perspective*, 2, 135-170.
- Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed). (pp. 17–64), Westport, CT: American Council on Education and Praeger.
- Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50, 1-73.
- Kang, T., Gutierrez Arvizu, M. N., Chaipupae, P., & Lesnov, R. (2016). Reviews of academic English listening tests for non-native speakers. *International Journal of Listening*. Published online on June 27. Retrieved from <https://www.tandfonline.com/doi/full/10.1080/10904018.2016.1185210>

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

- Kelly, P. (1991). Lexical ignorance: The main obstacle to listening comprehension with advanced foreign language learners, *IRAL*, 2, 135-149.
- Kintsch, W. (1998). *Comprehension*. Cambridge, UK: Cambridge University Press.
- Kline, P. (1993). *The handbook of psychological testing*. London, UK: Routledge.
- Kress, G. (2010). *Multimodality. A social semiotic approach to contemporary communication*. London, UK: Routledge.
- Kress, G., Jewitt, C., Bourne, J., Franks, A., Hardcastle, J., Jones, K., & Reid, E. (2005). *English in modern classrooms. A multimodal perspective on teaching and learning*. London, UK: RoutledgeFalmer.
- Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London, UK: Longman.
- Latifi, M., Tavakoli, M., & A'lipour, J. (2013). Investigating the effect of video materials on testing foreign language learners' listening performance. *Middle-East Journal of Scientific Research*, 13, 1197-1201.
- Lee, S. & Lee, S. (2015). Effects of audio-visual aids on foreign language test anxiety, reading, and listening comprehension, and retention in EFL learners. *Perceptual and Motor Skills: Perception*, 120, 576-590.
- Lesnov, R. (2017). Using videos in ESL listening achievement tests: Effects on difficulty. *Eurasian Journal of Applied Linguistics*, 3, 67-91.
- Lewis, T. R. (1958). Listening. *Review of Educational Research*, 28, 89-95.
- Li, Z. (2013). The issues of construct definition and assessment authenticity in video-based listening comprehension tests: Using an argument-based validation approach. *International Journal of Language Studies*, 7, 61-82.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

- Linacre, M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, 7, 328.
- Linacre, M. (2012a). *Many-facet Rasch measurement: Facets tutorial 1/2012*. Retrieved from <http://www.winsteps.com/a/ftutorial2.pdf>
- Linacre, M. (2012b). *A user's guide to FACETS Rasch-model computer programs*. Retrieved from <https://pdfs.semanticscholar.org/7008/95d53c7e5bf837b0602b4aa2eb702038d629.pdf>
- Linacre, M. (2013). Reliability, separation and strata: Percentage of sample in each level. *Rasch Measurement Transactions*, 26, 1399.
- Linacre, M. (2017) Facets computer program for many-facet Rasch measurement (Version 3.80.0) [Computer Software]. Beaverton, Oregon: Winsteps.com
- Londe, Z. (2009). The effects of video media in English as a second language listening comprehension test. *Issues in Applied Linguistics*, 17, 41-50.
- Lynch, T. (2011). Academic listening in the 21st century: Reviewing a decade of research. *Journal of English for Academic Purposes*, 10, 79-88.
- Lyons, J. (1977). *Semantics*. Cambridge, UK: Cambridge University Press.
- Malamitsa, K., Kokkotas, P., & Kasoutas, M. (2008). Graph/Chart interpretation and reading comprehension as critical thinking skills. *Science Education International*, 19, 371-384.
- Mason, A. (1983). *Understanding academic lectures*. Englewood Cliffs, NJ: Prentice Hall.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

- Mason, A. (1994). By dint of: Student and lecturer perceptions of lecture comprehension strategies in first-term graduate study. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 199-218). New York, NY: Cambridge University Press.
- Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Erlbaum.
- Mayer, R. (2002). Multimedia learning. *Psychology of Learning and Motivation*, 41, 85-139.
- Mayer, R. (2005). Cognitive theory of multimedia learning. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 31-48). Cambridge, UK: Cambridge University Press.
- Mayer, R. (2009). *Multimedia Learning*. Cambridge, UK: Cambridge University Press.
- McDonough, J. (1978). *Listening to lectures*. Oxford, UK: Oxford University Press.
- McNamara, T. (1996). *Measuring second language performance*. Essex, UK: Addison Wesley Longman Ltd.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York, NY: Macmillan Publishing.
- Mislevy, R., Steinberg, L., & Almond, R. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3-62.
- Morell, T., Garcia, M., & Sanchez, I (2008). Multimodal strategies for effective academic presentation in English for non-native speakers. In R. Monroy & A. Sanchez (Eds.), *25 years of applied linguistics in Spain: milestones and challenges* (pp. 557-568). Murcia, Spain: Universidad de Murcia Editum.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

- Morley, C. (2007). Listening: Top down and bottom up. Retrieved from <http://www.teachingenglish.org.uk/think/articles/listening-top-down-bottom>.
- Morley, J. (2001). Aural comprehension Instruction: Principles and practices. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (3rd ed.) (pp. 69-85). Boston, MA: Heinle & Heinle.
- Murphy, D., & Candlin, C. (1979). Engineering lecture discourse and listening comprehension. *Practical Papers in English Language Education*, 2, 1-79.
- NCTE Framework for 21st Century Curriculum and Assessment. (2013, February). Retrieved from <http://www.ncte.org/governance/21stcenturyframework>
- Nunnally J., & Bernstein L. (1994). *Psychometric theory*. New York: McGraw-Hill Higher, INC.
- Ockey, G. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing*, 24, 517–537.
- Open Doors 2016 (2016). Retrieved from <https://www.iie.org/Research-and-Insights/Open-Doors/Open-Doors-2016-Media-Information>
- Osborne, J. (1996). Beyond constructivism. *Science Education*, 80, 53-82.
- Paivio, A. (1979). *Imagery and verbal processes*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Paivio, A. (1991). Dual coding theory: retrospect and current status. *Canadian Journal of Psychology*, 45, 255-287.
- Paivio, A. (2006). Dual coding theory and education. Draft chapter for the conference on “Pathways to Literacy Achievement for High Poverty Children,” The University

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

of Michigan School of Education, September 29-October 1, 2006. Retrieved from http://coral.ufsm.br/tielletcab/Apostilas/DCT_Paivio.pdf

Paivio, A., & Lambert, W. (1981). Dual coding and bilingual memory. *Journal of Verbal Learning & Verbal Behavior*, 20, 532-539.

Parry, T., & Meredith, R. (1984). Videotape vs. audiotape for listening comprehension tests: An experiment. *OMLTA Journal*, 47-53. Retrieved from <http://files.eric.ed.gov/fulltext/ED254107.pdf>

Piaget, J. (1980). The psychogenesis of knowledge and its epistemological significance. In M. Piatelli-Palmarini (Ed.), *Language and learning* (pp. 23-34). Cambridge, MA: Harvard University Press.

Picou, E., Ricketts, T., & Hornsby, B. (2011). Visual cues and listening effort: Individual variability. *Journal of Speech Language and Hearing Research*, 54, 1416-1430.

Powers, D. (1985). *A survey of academic demands related to listening skills* (Research Rep. No. 20). Princeton, NJ: Educational Testing Service.

Progosh, D. (1996). Using video for listening assessment: Opinions of test-takers. *TESL Canada Journal*, 14, 34-44.

Purnell, K., & Solman, R. (1991). The influence of technical illustrations on students' comprehension of geography. *Reading Research Quarterly*, 26, 277-299.

Pusey, K. & Lenz, K. (2014). Investigating the interaction of visual input, working memory, and listening comprehension. *Language Education in Asia*, 5, 66-80.

Rankin, P. T. (1928). The importance of listening ability. *The English Journal*, 17, 623-630.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Reliability and Separation to Measures (2017). Retrieved from

<http://www.winsteps.com/winman/reliability.htm>

Research Methods Knowledge Based (2006, October). Retrieved from

<https://www.socialresearchmethods.net/kb/destypes.php>

Richards, J. (1983). Listening comprehension: Approach, design, procedure. *TESOL Quarterly*, 17, 219-240.

Rivers, W. (1981). *Teaching foreign language skills* (2nd ed.) Chicago, IL: University of Chicago Press.

Rosenfeld, M., Leung, S., & Oltman P. (2001). *The Reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels* (TOEFL Monograph No. MS-21). Princeton, JN: Educational Testing Service.

Rost, M. (2005). L2 listening. In Hinkel, E. (Ed.). *Handbook of research in second language teaching and learning* (pp. 503-527). Mahwah, NJ: Lawrence Earlbaum Associates, Publishers.

Rost, M. (2016). *Teaching and researching: Listening* (3rd ed.). New York, NY: Routledge.

Rowley-Jolivet, E. (2002). Visual discourse in scientific conference papers: a genre-based study. *English for Specific Purposes*, 21, 19–40.

Rubin, J. (1995). The contribution of video to the development of competence in listening. In D. Mendelsohn & J. Rubin (Eds.), *A guide for the teaching of second language listening* (pp. 151-165). San Diego: Dominic Press.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

- Schumacker, R., & Lomax, R. (2004). *A beginner's guide to structural equation modeling*. (2nd Ed.) Mahwah, NJ: Lawrence Erlbaum Associates.
- Sébille, V., Blanchin, M., Guillemin, F., Falissard, B., & Hardouin, J.-B. (2014). A simple ratio-based approach for power and sample size determination for 2-group comparison using Rasch models. Retrieved from <https://bmcmmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-14-87>
- Shin, D. (1998). Using videotaped lectures for testing academic listening proficiency. *International Journal of Listening*, 12, 57–80.
- Shohamy, E. & Inbar, O. (1991). Construct validation of listening comprehension tests: The effect of text and question type. *Language Testing*, 8, 23-40.
- Sternberg, R. J. (2003). *Cognitive theory* (3rd ed.). Belmont, CA: Thomson Wadsworth.
- Sueyoshi, A., & Hardison, D. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, 55, 661-699.
- SurveyGizmo: Professional survey solution. (2017) [Computer Software]. Available from <https://www.surveygizmo.com/>
- Suvorov, R. (2009). Context visuals in L2 listening tests: The effects of photographs and video vs. audio-only format. In C. Chapelle, H. Jun, & I. Katz (Eds.), *Developing and evaluating language learning materials* (pp. 53-68). Ames, IA: Iowa State University.
- Suvorov, R. (2013). *Interacting with visuals in L2 listening tests: An eye-tracking study*. (Doctoral dissertation). Retrieved from <https://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=4306&context=etd>

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

- Suvorov, R. (2015a). The use of eye tracking in research on video-based second language (L2) listening assessment: A comparison of context videos and content videos. *Language Testing*, 21, 1-21.
- Suvorov, R. (2015b). Interacting with visuals in L2 listening tests: An eye-tracking study. In V. Berry (Ed.), *ARAGs research reports online (Report #AR-A/2015/1)*. Retrieved from British Council
http://www.britishcouncil.org/sites/default/files/interacting_with_visuals_in_l2_listening_tests_suvorov.pdf
- Tabachnick, B., & Fidell, L. (2013). *Using multivariate statistics* (6th ed). Boston, MA: Pearson.
- Tafaghodtari, M.H., & Vandergrift, L. (2008). Second and foreign language listening: Unraveling the construct. *Perceptual and Motor Skills*, 107, 99-113.
- Tate, R. (1984). Limitations of centering for interactive models. *Sociological Methods & Research*, 13, 251–271.
- The Academic Word List (2017). Retrieved from
<http://www.oxfordlearnersdictionaries.com/about/academic>
- The Oxford Text Checker (2017). Retrieved from
http://www.oxfordlearnersdictionaries.com/oxford_3000_profiler
- Toulmin, S. (2003). *The uses of argument* (updated edition). Cambridge, UK: Cambridge University Press.
- Tucker, M. (1991). A compendium of textbook views on planned versus post hoc tests. In B. Thompson (Ed.), *Advances in education research: Substantive findings, methodological developments* (Vol. 1, pp. 107-118). Greenwich, CT: JAI Press.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

University of Delaware (2015, February 11). *Security expert Michael Chertoff discusses cybersecurity challenges, solutions* [Video File]. Available from

<https://www.youtube.com/watch?v=3MkFO6EALI8>

Ur, P. (1984). *Teaching listening comprehension*. Cambridge, UK: Cambridge University Press.

van Dijk, T., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York, NY: Academic Press.

Vandergrift, L. (2004). Listening to learn or learning to listen? *ARAL*, 24, 3-25.

Vandergrift, L. (2005). Relationships among motivation orientations, metacognitive awareness and proficiency in L2 listening. *Applied Linguistics*, 26, 70-89.

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching*, 40, 191-210.

Vandergrift, L., & Goh, C. (2012). *Teaching and learning second language listening: Metacognition in action*. New York, NY: Routledge.

Visual (2017). In *CollinsDictionary.com*. Retrieved from

<https://www.collinsdictionary.com/dictionary/english/visual>

Vygotsky, L. S. (1962/1986). *Thought and language*. Cambridge, MA: MIT Press.

Wagner, E. (2007). Are they watching? Test-taker viewing behavior during an L2 video listening test. *Language Learning & Technology*, 11, 67-86.

Wagner, E. (2008). Video listening tests: What are they measuring? *Language Assessment Quarterly*, 5, 218-243.

Wagner, E. (2010a). Test-takers' interaction with an L2 video listening test. *System*, 38, 280-291.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

- Wagner, E. (2010b). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing*, 27, 493-513.
- Wagner, E. (2013). An investigation of how the channel of input and access to test questions affect L2 listening test performance. *Language Assessment Quarterly*, 10, 178-195.
- Wagner, E., & Schönau, D. (Eds.) (2016). *Common European framework of reference for visual literacy*. New York, NY: Waxman.
- Wainer, H. & Keily, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, 24, 185-201.
- Walma van der Molen, J. (2001). Assessing text-picture correspondence in television news: The development of a new coding scheme. *Journal of Broadcasting and Electronic Media*, 45, 483–498.
- Wetzel, C., Radtke, P., & Stern, H. (1994). *Instructional effectiveness of video media*. Hillsdale, NJ: Lawrence Erlbaum.
- Williams, M. (2016, November 11). *Psyc123 Lec 18: The Issues, the fights and who controls the frame*. [Video File]. Available from <https://www.youtube.com/watch?v=isUy2dKkJ0s>
- Wolff, D. (1987). Some assumptions about second language text comprehension. *Studies in Second Language Acquisition*, 9, 307-326.
- Woods, N. (1978). *College reading and study skills* (3rd ed.). New York, NY: Holt, Rinehart, and Winston.
- Wright, B. (1992). Raw Scores Are Not Linear Measures: Rasch vs. Classical Test Theory CTT Comparison. *Rasch Measurement Transactions*, 6, 208.

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

YaleCourses (2008, November 18). *What is biomedical engineering?* [Video File].

Available from <https://www.youtube.com/watch?v=mRu9SDMWn2g>

YaleCourses (2013, June 2). *Human Emotion 10.2: Emotions in a Social World II (Social*

Emotions). [Video File]. Available from

<https://www.youtube.com/watch?v=L7W4wpkfd1Y>

Yang, H-Y. (2014). Does multimedia support individual differences? EFL learners'

listening comprehension and cognitive load. *Australasian Journal of Educational*

Technology, 30, 699-713.

Young, L. (1994). University lectures – macro-structure and micro-structure. In J.

Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 159-176). New

York, NY: Cambridge University Press.

Appendix A1

Sampling Frame for TESOL-Affiliated Organizations

Asia and Oceania

1. Australian Council of TESOL Associations (ACTA)
2. Bangladesh English Language Teachers Association (BELTA)
3. English Language Teachers' Association of India (ELTAI)
4. English Language Teachers' Association of Mongolia (Mongolia TESOL)
5. English Teachers' Association, Taiwan
6. Hong Kong Association for Applied Linguistics (HAAL)
7. Forum for Teachers of English Language and Literature, India (Fortell)
8. Japan Association for Language Teaching (JALT)
9. Korea TESOL (KOTESOL), South Korea
10. Nepal English Language Teachers' Association (NELTA)
11. Penang English Language Learning and Teaching Association (PELLTA), Malaysia
12. Philippine Association for Language Teaching (PALT)
13. Society of Pakistani English Language Teachers (SPELT)
14. Teachers of English as a Foreign Language in Indonesia (TEFLIN)
15. TESOL Association of Aotearoa New Zealand (TESOLANZ)
16. Thailand TESOL (ThaiTESOL)

Europe and Eurasia

17. Association of Teachers of English in the Czech Republic (ATE-CR)
18. Associacao Portuguesa de Professores de Ingles (APPI), Portugal
19. Azerbaijan English Teachers' Association (AzETA)
20. Bulgarian English Teachers' Association (BETA)
21. Center for English Teaching Excellence (CETE)
22. Georgia Croatian Association of Teachers of English (HUPE)
23. English Teachers' Association of Georgia (ETAG), Georgia
24. English Language Teachers' Association, Albania
25. English Language Teachers' Association of Serbia (ELTA)
26. English Language Teachers' Association of Macedonia (ELTAM)
27. IATEFL Poland
28. Moldova English Teachers' Association (META)
29. National Association of Language Development in the Curriculum (NALDIC), England
30. National Association of Teachers of English in Russia (NATE Russia), Moscow, Russia
31. TESOL France
32. TESOL Greece
33. TESOL Italy
34. TESOL Ukraine
35. TESOL-Spain
36. Yakut TESOL (YAKTESOL), Yakut, Russia

Caribbean, Central and South America

37. Argentina TESOL (ARTESOL)
38. Asociacion Colombiana de Profesores de Ingles (ASOCOPI), Colombia
39. Asociacion Costarricense de Profesores de Ingles (ACPI), Costa Rica
40. Bolivian English Teachers Association (BETA)
41. Brazil TESOL (BRAZ-TESOL)
42. Dominican Republic TESOL
43. Federacion Nacional de Profesores de Ingles de Universidades y Politecnicas del Ecuador (FENAPIUPE), Ecuador
44. Grupe de Especialistas En Lengua Inglesa (GELI), Cuba
45. Honduran English Language Teachers Association (HELTA)

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

46. Miragoane Association of Teachers of English (MATE), Haiti
47. Nicaraguan English Language Teachers' Association (ANPI)
48. Panama TESOL
49. Peru TESOL Association
50. Puerto Rico TESOL (PRTESOL)
51. TESOL Chile
52. Uruguay TESOL (URUTESOL)
53. Venezuela TESOL (VENTESOL)

Africa and the Middle East

54. Addis Ababa English Language Teachers' Association, Ethiopia
55. Association of Teachers of English in Senegal (ATES)
56. Burkina English Teachers Association (BETA)
57. Cameroon English Language and Literature Teachers Association (CAMELTA)
58. English Language Education Association of Turkey (ELEA-Inged)
59. English Teachers Association of Israel (ETAI)
60. Kingdom of Saudi Arabia Association of Language Teachers (KSAALT)
61. Libya TESOL
62. Malian Association of Teachers of English
63. Moroccan Association of Teachers of English
64. Nile TESOL, Egypt
65. Qatar TESOL
66. Tanzanian English Language Teachers' Development Meeting (TELTDMD)
67. TESOL Arabia, United Arab Emirates
68. TESOL Kuwait
69. TESOL Sudan
70. Tunisia TESOL

North America

71. Alabama-Mississippi TESOL (AMTESOL), USA
72. Alaska Association of Bilingual Education (AKABE), USA
73. Arizona TESOL (AZTESOL), USA
74. Arkansas TESOL (ARKTESOL), USA
75. Asociación Nacional Universitaria de Profesores de Inglés (ANUPI-TESOL), Mexico
76. British Columbia Teachers of English as an Additional Language (BC TEAL)
77. California and Nevada TESOL (CATESOL), USA
78. Carolinas TESOL (Carolina TESOL), North & South Carolina, USA
79. Colorado TESOL (CoTESOL), USA
80. Connecticut TESOL (ConnTESOL), USA
81. Dakota TESL
82. Georgia TESOL (GATESOL), USA
83. Hawaii TESOL, USA
84. Illinois TESOL/BE (ITBE), USA
85. Indiana TESOL (INTESOL), USA
86. Intermountain TESOL (ITESOL), Utah, Idaho, and Wyoming, USA
87. Kentucky TESOL (KYTESOL), USA
88. Louisiana TESOL (LaTESOL), USA
89. Maryland TESOL (MDTESOL), USA
90. Mexican Assn. of English Teachers (MEXTESOL), Mexico
91. Massachusetts TESOL (MATESOL)
92. Michigan TESOL (MITESOL), USA
93. Mid-America TESOL (MIDTESOL), Kansas, Iowa, Nebraska, Missouri, USA
94. Minnesota TESOL (MinneTESOL), USA
95. New Jersey TESOL/New Jersey Bilingual Educators (NJTESOL/NJBE), USA
96. New Mexico TESOL (NMTESOL), USA
97. New York State TESOL (NYSTESOL), USA

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

98. Northern New England TESOL (NNETESOL), Maine, New Hampshire, and Vermont, USA
99. Ohio TESOL, USA
100. Oklahoma TESOL (OKTESOL), USA
101. Oregon TESOL (ORTESOL), USA
102. PennTESOL-East, Pennsylvania (Eastern), South New Jersey, & Delaware, USA
103. Rhode Island Teachers of English Language Learners (RITELL), USA
104. Sunshine State TESOL (SSTESOL), Florida, USA
105. Tennessee TESOL (TNTESOL), USA
106. TESL Association of Ontario (TESL Ontario), Ontario, Canada
107. TESL Nova Scotia, Nova Scotia, Canada
108. Teachers of English as a Second Language of New Brunswick (TESL NB)
109. TEXTESOL-II, San Antonio, Texas, USA
110. TEXTESOL-III, Austin, Texas, USA
111. TEXTESOL-IV, Houston, Texas, USA
112. TEXTESOL-V, Dallas, Texas, USA
113. Three Rivers TESOL (3-R TESOL), Western Penn. and West Virginia, USA
114. Virginia TESOL (VATESOL), USA
115. Washington Area TESOL (WATESOL), Washington, D.C., USA
116. Washington Association for the Education of Speakers of Other Languages (WAESOL),
Washington, USA
117. West Virginia TESOL (WVTESOL), USA
118. Wisconsin TESOL (WITESOL), USA

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Appendix A2

TESOL Affiliates Selected for the Study

TESOL affiliate	URL link	Contact information
Asia and Oceania		
Australian Council of TESOL Associations (ACTA)	http://www.tesol.org.au/	president@tesol.org.au (Michael Michell, president) secretary@tesol.org.au (Margaret Turnbull, secretary)
Bangladesh English Language Teachers Association (BELTA)	http://www.belta-bd.org/	info@belta-bd.org
Japan Association for Language Teaching (JALT)	https://jalt.org/	https://jalt.org/contact/form
Korea TESOL (KOTESOL), South Republic of Korea	https://koreatesol.org/	https://koreatesol.org/contact
Nepal English Language Teachers' Association (NELTA)	http://www.nelta.org.np/	http://www.nelta.org.np/contact cnelta@gmail.com
Penang English Language Learning and Teaching Association (PELLTA)	http://www.pellta.org/	pelltapenang@gmail.com
Philippine Association for Language Teaching (PALT)	https://paltphilipinas.wordpress.com/author/paltphilipinas/	paltphil@gmail.com (Ms. Marge C. Ballesteros, international liaison officer)
Europe and Eurasia		
Association of Teachers of English in the Czech Republic (ATE-CR)	http://atecr.weebly.com/	atecr@centrum.cz (Ms. Libuše Kohutová, president) ivahavlikova@email.cz (Ms. Iva Havlíková, membership secretary) http://atecr.weebly.com/contacts.html
Azerbaijan English Teachers' Association (AzETA)	http://www.az-eta.org/	http://www.az-eta.org/contact.php
English Teachers' Association of Georgia (ETAG)	https://www.facebook.com/ETAG-in-Georgia-152713621457305/ http://www.etag.ge/	Central Office, Tbilisi tsisanat@yahoo.com (Tsisana Tsiskaridze) lalimdi@yahoo.com (Lali Mdinardze) etag.tbilisi@caucasus.net etag.courses@caucasus.net
Moldova English Teachers' Association (META)	http://meta-moldova.md/	admin@meta-moldova.md
Portugal: Associação Portuguesa de Professores de Inglês (APPI)	http://www.appi.pt/	socios@appi.pt appi@appi.pt
United Kingdom: National Association of Language Development in the Curriculum (NALDIC)	https://naldic.org.uk/	enquiries@naldic.org.uk
TESOL France	https://www.tesol-france.org/en/	https://www.tesol-france.org/en/contact.html
TESOL Italy	http://tesolitaly.org/new/	http://tesolitaly.org/new/contact-us/
National Association of Teachers of English in Russia (NATE Russia)	http://nate-russia.ru/	http://nate-russia.ru/contacts.php
Caribbean, Central, and South America		
The Colombian Association of Teachers of English (ASOCOPI)	http://www.asocopi.org/en/inicio.html	asocopicolombia@gmail.com
Brazil TESOL (BRAZ-TESOL)	http://www.braztesol.org.br/site/view.asp	http://www.braztesol.org.br/site/view.asp?p=8 info@braztesol.org.br

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Panama TESOL	http://www.panamatesol.org/	info@panamatesol.org fernandodeleontesol@gmail.com (Fernando De León) http://www.panamatesol.org/contact_us
Peru TESOL Association	http://www.perutesol.org/	info@perutesol.org http://www.perutesol.org/contact.php
Puerto Rico TESOL	http://www.prtesol.org/	tesolpuertorico@gmail.com http://www.prtesol.org/contact-us.html
TESOL Chile	http://tesolchile.cl/welcome/	http://tesolchile.cl/welcome/contact-2/
Uruguay TESOL (URUTESOL)	http://urutesol.org/	urutesol@gmail.com https://www.facebook.com/urutesol
<hr/>		
Africa and the Middle East		
Cameroon English Language and Literature Teachers Association (CAMELTA)	http://camelta-cameroon.weebly.com/index.html	hkuchah@yahoo.com
English Language Education Association of Turkey (INGED)	http://www.inged.org.tr/	http://www.inged.org.tr/index.php?option=com_contact&view=contact&id=2&Itemid=65
Qatar TESOL	http://qatartesol.org/	https://www.facebook.com/Qatar-Tesol-112664268821155/
TESOL Arabia	http://www.tesolarabia.co/	info@tesolarabia.org http://www.tesolarabia.co/contact-us/
TESOL Kuwait	http://www.tesolkuwait.org/	president@tesolkuwait.com (president) Secretary@tesolKuwait.org (secretary) Others at http://www.tesolkuwait.org/contact-us.html
TESOL Sudan Moroccan Association of Teachers of English	http://www.tesolsudan.net/ http://mate.ma/	hindmoelyas@yahoo.com http://mate.ma/index.php/contact
<hr/>		
North America		
Alabama-Mississippi TESOL (AMTESOL)	https://www.amtesol.org/	AStamps@international.msstate.edu (president) diamoms@auburn.edu (secretary)
Arkansas TESOL (ARKTESOL)	http://www.arktesol.org/	tricia.kerr@arkansas.gov (president) cjay@bentonvillek12.org (secretary)
Colorado TESOL (CoTESOL)	http://www.colorado.edu/iec/cotesol/	Larry.Fisher@colorado.edu
Dakota TESL	http://dakotatesl.com/	whiplk1@ndseec.com (president) heather.glidewell@lsssd.org (secretary)
Hawaii TESOL	http://hawaiitesol.wildapricot.org/	mark.wolfersberger@byuh.edu
Indiana TESOL (INTESOL)	http://www.intesol.org/	williamsonnt@gmail.com (president)
Intermountain TESOL (ITESOL), Utah, Idaho, and Wyoming	http://itesol.org/	ITESOL Google Group: https://groups.google.com/forum/#!forum/itesol
Maryland TESOL (MDTESOL)	https://www.mdtesol.org/	website@mdtesol.org
Massachusetts TESOL (MATSOL)	http://www.matsol.org/	matsol@matsol.org
Minnesota TESOL (MinneTESOL)	http://minnetesol.org/	admin@minnetesol.org
New Jersey TESOL/New Jersey Bilingual Educators (NJTESOL/NJBE)	http://www.njtesol-njbe.org/	webmaster@njtesol-njbe.org
New York State TESOL (NYSTESOL)	http://www.nystesol.org/	membershipinquiries@nystesol.org

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

PennTESOL-East, Pennsylvania (Eastern), South New Jersey, & Delaware	http://www.penntesol-east.org/	pennsylvaniaatesoleast@gmail.com
Sunshine State TESOL (SSTESOL), Florida	http://sstesol.org/	http://sstesol.org/?page_id=804
TESL Association of Ontario (TESL Ontario)	http://www.teslontario.org/	http://www.teslontario.org/staff
TEXTESOL-II, San Antonio, Texas	textesoltwo.org	https://www.facebook.com/TEXTESOL/
TEXTESOL-III, Austin, Texas	http://www.textesol3.org/	president@textesol3.org (president) secretary@textesol3.org (secretary)
Three Rivers TESOL (3-R TESOL), Western Penn. and West Virginia	https://threeriverstesol.org/wp/	president@threeriverstesol.org (president) secretary@threeriverstesol.org (secretary)
Washington Area TESOL (WATESOL), Washington, D.C.	https://watesol.org/	watesolmembership@gmail.com
Wisconsin TESOL (WITESOL)	http://witesol.com/	WITESOL.President@gmail.com (president) witesolboard@gmail.com

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Appendix B

Video Listening Passages Found on the Internet

Title (Citation)	Content-related visual cues	Lecture type	Link / Time boundaries / University affiliation
Homeostasis (YaleCourses, 2008)	Visuals not accessible	traditional	https://www.youtube.com/watch?v=mRu9SDMWn2g 29:10 – 34:38 Yale University, USA
Food Tax (Williams, 2016)	Graphs, diagrams, pictures, text	traditional	https://www.youtube.com/watch?v=isUy2dKkJOs 1:07:34 – 1:14:36 Yale University, USA
Compassion (YaleCourses, 2013)	Graphs, schemes, pictures, text	online	https://www.youtube.com/watch?v=L7W4wpkfd1Y 13:01 – 18:20 Yale University, USA
Exoplanets (ASTR105x - Alien Worlds, 2014)	Diagrams, pictures, text	online	https://www.youtube.com/watch?v=cYmBoAEQtQI 01:10 – 07:10 Boston University, USA

Appendix C1

Recording Instructions

Prior to recording: Familiarizing with the script

1. Read the script attentively.
2. Watch the lecture from which the script was derived (see the link in the email).
3. Read the script again.
4. Read the outline for the script. It consists of outlines for each paragraph.
5. Rehearse your lecture speech, using the outlines sparingly.
6. Repeat steps 2 through 5 if needed.

Recording

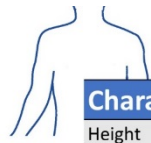
1. Closely read the first paragraph of the script.
2. Read the outline for the first paragraph of the script.
3. Deliver the content of the paragraph orally, by memory. You can occasionally use the outline.
4. Repeat steps 1 through 3 for each of the paragraphs in the script.

Appendix C2

Visual Configurations of the Four ALC Test Videos

Homeostasis video

1



Characteristic	Average Value
Height	170 cm
Weight	60-80 kgs
Surface area	1.9 sq. m
Temperature	37°C

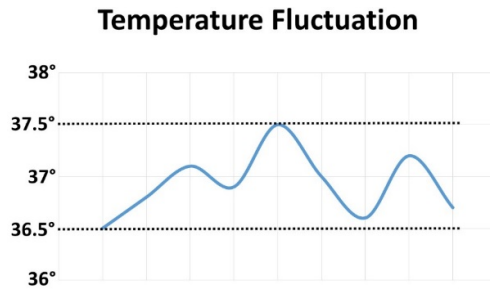
- a. graph
- b. illustrating, supplementing
- c. word count: 12

2



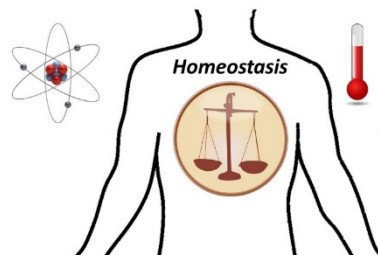
- a. picture
- b. illustrating
- c. no text

3



- a. graph
- b. illustrating, organizing
- c. word count: 2

4



- a. picture
- b. illustrating
- c. word count: 1

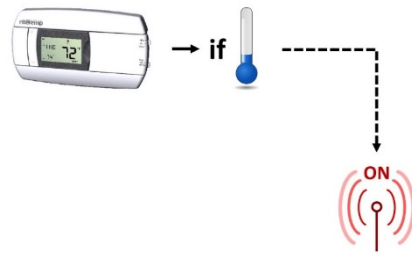
CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

5



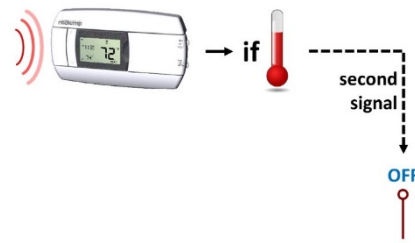
- a. picture
- b. illustrating
- c. no text

6



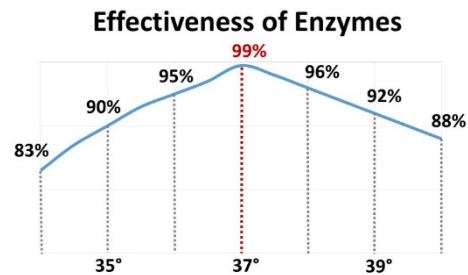
- a. graph
- b. illustrating, organizing
- c. word count: 2

7



- a. graph
- b. illustrating, organizing
- c. word count: 4

8



- a. graph
- b. illustrating, organizing, supplementing
- c. word count: 3

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

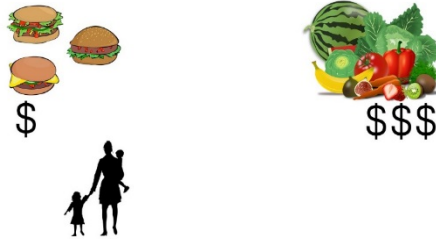
Food Tax video

1



- a. picture
- b. illustrating
- c. word count: 1

2



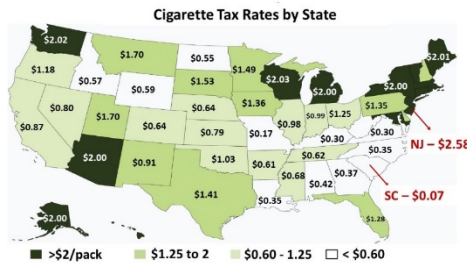
- a. picture
- b. illustrating
- c. no text

3



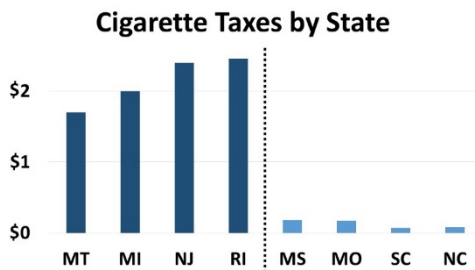
- a. picture
- b. illustrating
- c. no text

4



- a. graph
- b. illustrating, supplementing
- c. word count: 8

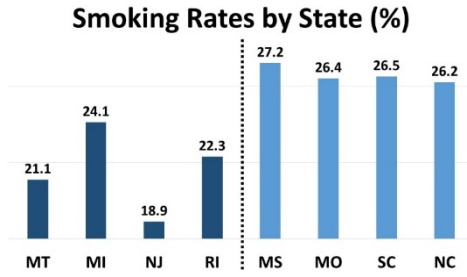
5



- a. graph
- b. illustrating, supplementing
- c. word count: 11

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

6



- a. graph
- b. illustrating, supplementing
- c. word count: 11

7



- a. graph
- b. illustrating, organizing
- c. word count: 1

8



- a. picture
- b. illustrating
- c. no text

CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

Compassion video

1



- a. picture
- b. illustrating
- c. no text

2



- a. picture
- b. illustrating
- c. no text

3



- a. picture
- b. illustrating
- c. no text

4



- a. picture
- b. illustrating
- c. no text

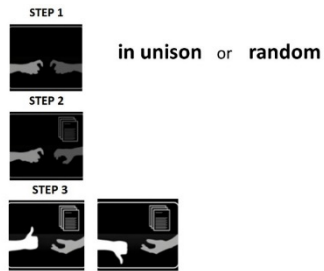
5



- a. picture
- b. illustrating
- c. no text

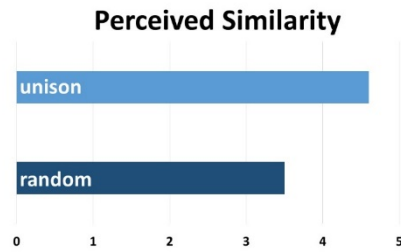
CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

6



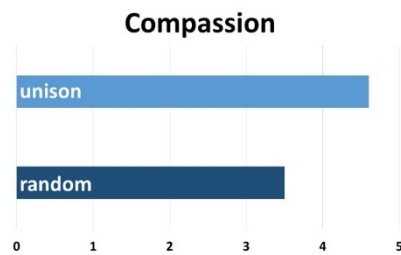
- a. graph
- b. illustrating, organizing
- c. word count: 7

7



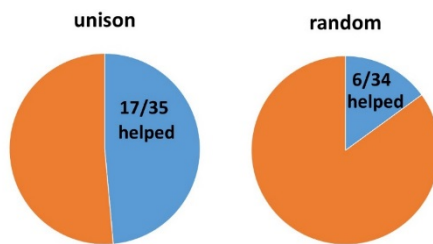
- a. graph
- b. illustrating
- c. word count: 4

8



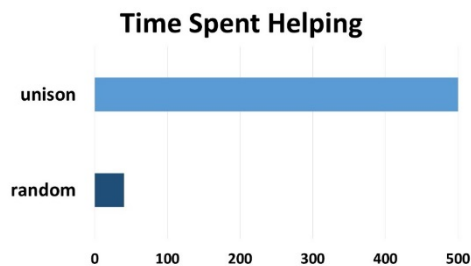
- a. graph
- b. illustrating
- c. word count: 3

9



- a. graph
- b. illustrating
- c. word count: 4

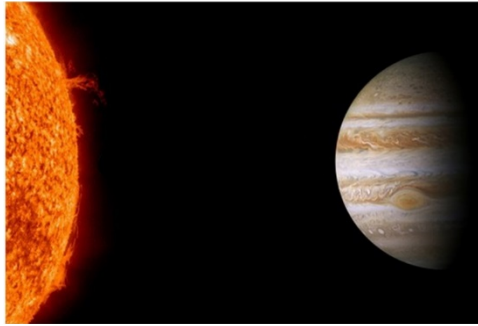
10



- a. graph
- b. illustrating
- c. word count: 5

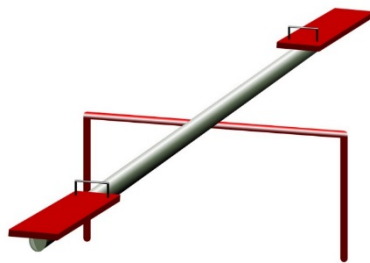
Exoplanets video

1



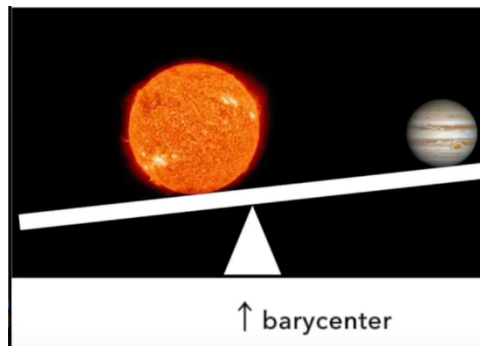
- a. picture
- b. illustrating
- c. no text

2



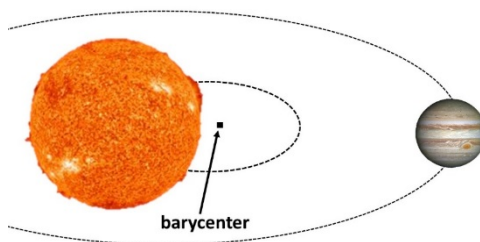
- a. picture
- b. illustrating
- c. no text

3



- a. picture
- b. illustrating
- c. word count: 1

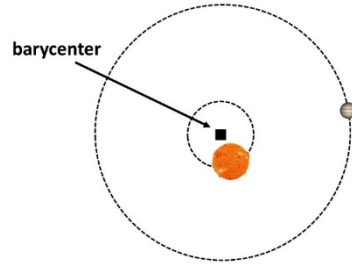
4



- a. graph
- b. illustrating,
organizing
- c. word count: 1

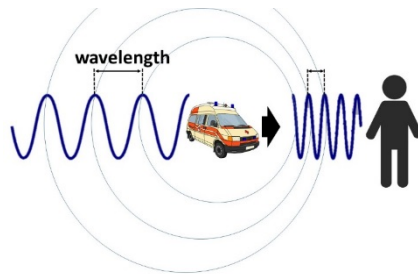
CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

5



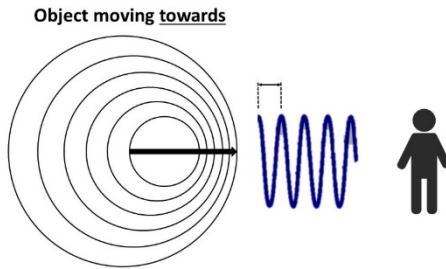
- a. graph
- b. illustrating, organizing
- c. word count: 1

6



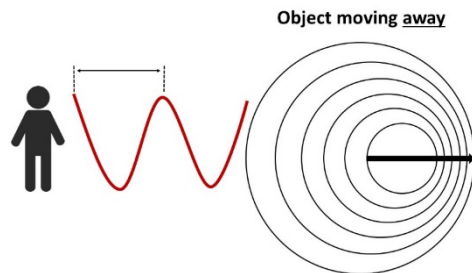
- a. graph
- b. illustrating, organizing
- c. word count: 1

7



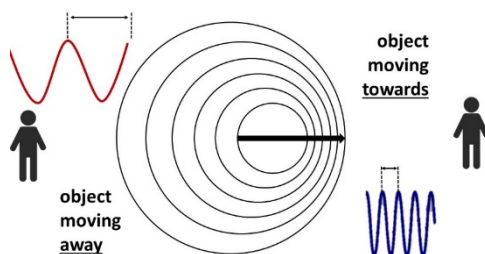
- a. graph
- b. illustrating, organizing
- c. word count: 3

8



- a. graph
- b. illustrating, organizing
- c. word count: 3

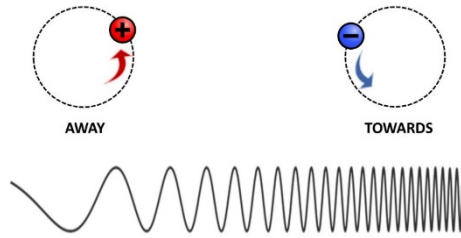
9



- a. graph
- b. illustrating, organizing
- c. word count: 6

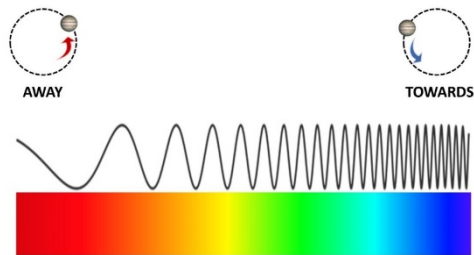
CONTENT-RICH VISUALS IN L2 ACADEMIC LISTENING CONSTRUCT

10



- a. graph
- b. illustrating
- c. word count: 2

11



- a. graph
- b. illustrating
- c. word count: 2

Appendix D

ALC Test: Consent, scripts, items, specifications

Informed Consent

You are invited to participate in a research study titled “**The Role of Content-Rich Videos in the L2 Academic Listening Assessment Construct**” This study is being done by Roman Lesnov from Northern Arizona University.

The purpose of this research study *is to justify the use of visual information in second language tests of listening comprehension*. If you agree to take part in this study, you will be asked to complete an online listening test and brief online questionnaires. For the test, you will listen to several lectures and answer comprehension questions. For the questionnaires, you will judge how helpful visual information is for your listening. It will take you approximately **40 minutes** to complete.

Your name will be drawn into a raffle to win one of 40\$ prizes. Participation in the research is not required in order to participate in the drawing. We hope that your participation in the study will help shape the future of second language listening tests. We believe there are no known risks associated with this research study; however, as with any online related activity the risk of a breach of confidentiality is always possible. To the best of our ability your answers in this study will remain confidential. We will minimize any risks by **maintaining the data confidentially and securely**.

Your participation in this study is completely voluntary and you can withdraw at any time. You are free to skip any question that you choose. If you choose not to participate it will not affect your relationship with Northern Arizona University or result in any other penalty or less of benefits to which you are otherwise entitled.

If you have questions about this project or if you have a research-related problem, you may contact the researcher(s), **Roman Lesnov at (+1) 929-225-9330**. If you have any questions concerning your rights as a research subject, you may contact Northern Arizona University IRB Office at irb@nau.edu or (928) 523-9551.

By clicking “NEXT”, I affirm that I am over 18 years of age and agree that the information may be used in the research project described above.

Testlet 1. Homeostasis (Questions 1-6)

I want to talk about some concepts in physiology that are really important for this course in biomedical engineering. I want you to try to imagine a table that has characteristics of an average person – an adult male, 30 years old, average height, average weight, average surface area, ah average temperature, just a lot of average characteristics of an average person. And let's just take a look at one of these, let's look at weight. So weight is something that is actually a very carefully controlled parameter for a person. Ahm we take in a lot of food, we take in a lot of drink ah but we don't really gain a lot of weight, our weight stays pretty stable. And if you try to lose weight - you're too young to try to lose weight too much, but as you get older your metabolism changes, you realize how hard it is to lose weight, and we know it's hard because we spend so much energy talking about it. Now ah weight is pretty carefully controlled and your body does it on its own, you don't have to think about it. Now ah also, temperature. Temperature is something that is within a narrow range, stays pretty constant. You go from inside to outside, you go into a hot room, your temperature doesn't change that much, it stays within this range of 36.5 to 37.5 degrees. And it's so stable, it's so important that it's stable that when it changes just a little bit, we know that something is wrong. You measure your temperature, it goes up and down. And if it's a little bit up, we know something's wrong – you have a fever. We know it because it's so stable.

So, you could go through a lot of these parameters and think about them in the same way that these things are really very highly controlled. And this process of control to maintain a constant environment within our bodies, whether it's mass or chemical composition, or temperature, is called homeostasis. And your body has very elaborate mechanisms for maintaining this state of homeostasis. Ah in spite of the fact that we take in a lot of chemicals and ah in different ways, and we have to do that to stay alive, but we have mechanisms to control the process very well. Now homeostasis is enabled by both complex and simple control mechanisms. And we can describe them in ways that are actually probably pretty similar to control mechanisms mechanisms that you're already familiar with. So, let's take for example the thermostat in your dorm. Maybe this is a bad example, maybe you don't have control over your thermostat or maybe your thermostat doesn't work very well. But just imagine a perfect thermostat. No matter what the

CONTENT-RICH VIDEOS IN L2 ACADEMIC LISTENING CONSTRUCT

temperature it is outside, it maintains the constant temperature inside your room. Now this perfect thermostat works through a control mechanism that's called negative feedback. And so it works like this. You have a thermostat that's measuring the temperature and it's sending signals to a heater somewhere. And when the temperature level drops below a certain level, then it sends a signal to turn on, the heater turns on, and it's just heating, it's just heating until it receives the second signal. So when does it receive the second signal? When the temperature goes above the certain level, then the second signal is sent, and it turns off. So the heater's on, it's just heating, heating, heating and it gets the signal to turn off. It says 'oh we've gone too high', and it shuts down. So our bodies have these same mechanisms like that, they mainly use this principle of negative feedback to control the parameters that are important for life within certain ranges.

So why is temperature, for example, so important to keep at 37 degrees? Well it's because that's the temperature at which many of the molecules in our bodies operate most efficiently. So enzymes are the best example of this. Enzymes are molecules that catalyze chemical reactions and our bodies are basically networks of chemical reactions, and enzymes operate most effectively at 37 degrees Celsius. So when we're off from that temperature then enzymes don't work properly any more, and then the chemical reactions don't run as well as they should. And there are other examples as well, but that's why it's important.

CONTENT-RICH VIDEOS IN L2 ACADEMIC LISTENING CONSTRUCT

1. According to the speaker, which statement about weight is true?
 - (A) Young people now have serious weight issues.
 - (B) We do not have to help our body control its weight.
 - (C) Eating food makes our weight quite unstable.
 - (D) We do not normally talk much about weight.

2. The normal body temperature range is _____ degrees Celsius.
 - (A) 36.5-37.0
 - (B) 36.0-37.5
 - (C) 36.5-37.5
 - (D) 37.5-38.5

3. We can infer that thermostats are _____.
 - (A) quite familiar to students
 - (B) weakly related to the lecture
 - (C) not helpful for understanding homeostasis
 - (D) in a perfect condition in college dorm rooms

4. Our body will most likely send the second control signal when _____.
 - (A) we have a fever
 - (B) we are cold
 - (C) our temperature is normal
 - (D) our temperature drops fast

5. Temperature control is important because it _____.
 - (A) slows harmful chemical reactions
 - (B) helps molecules work effectively
 - (C) increases the number of enzymes
 - (D) manages the body's feedback

6. This lecture is mainly about _____.
 - (A) how our body keeps its weight constant
 - (B) which body parameters are most important
 - (C) why body temperature is important
 - (D) how our body controls its environment

Testlet 2. Food Tax (Questions 7-12)

So today let's return to that idea of unhealthy foods that we've been talking about and think about how it interacts with taxes. Now, the most radical change of all when it comes to proposed ahm policies and food politics has to do with the idea of taxes. Taxing foods and will it actually be viable to put a tax on certain foods to help improve public health? And the rationale for doing something like this with taxes has to do with what we've been talking about in class. Those un ah healthy foods just simply cost more to make and to provide than unhealthy foods do. As a result, those unhealthy foods are more affordable for the poor. We could you use a tax policy to to discourage that affordability of unhealthy foods and we could take that money and use it as a subsidy for the foods we want, fresh produces, fruits and vegetables. And this is a topic that we've been thinking about for years. There is a precedent for this in the arena of tobacco. Now you know there're different taxes on packs of cigarettes that vary state by state by state around the country. And there's a huge difference between the biggest taxes of about two dollars and fifty cents a pack in New Jersey and Rhode Island versus the smallest tax of ah seven cents a pack in South Carolina. And the research in this areas has shown for years that taxes are the single most effective way to curb smoking. Other things do matter but taxes are the most effective. Those are current data that I just presented. But I also have data that are about a year older. If you compare the four states with the highest tax and the four with the lowest. So that's Montana, Michigan, New Jersey, and Rhode Island ah more than two dollars a pack, and Mississippi, Missouri, and the Carolinas, less than 20 cents a pack. You can see that difference is huge.

Of course, you can probably guess what I'm going to tell you next, which is the rate of smoking in the state with higher versus lower tobacco taxes. There's not a perfect relationship because in Michigan we can see quite a high level of smoking despite having one of the highest taxes in the country. But in general, ahm we can see that states with higher taxes have remarkably lower rates of smoking; the states with low cigarette taxes do have many more smokers. So taxes do matter, they do affect behavior. And we wonder if there could be something equivalent in the area of food. To show you just how much of a difference ah these taxes can make, let's look at California. In California, there

CONTENT-RICH VIDEOS IN L2 ACADEMIC LISTENING CONSTRUCT

is a heavy tax on cigarettes, with the money specifically earmarked to go to anti-tobacco programs, and that doesn't happen in every state. This act started in 1988 with a twenty-five cents per pack increase in taxes on cigarettes. And it generated about ninety million dollars a year, all going to these anti-tobacco act campaigns. And you might have seen those Truth Campaign ads that painted tobacco executives act really negatively. By 1999, this resulted in a twenty-seven percent decrease in smoking and nineteen percent decrease in deaths due to lung cancer, about 10 percent better than the rest of the country. Now that's a powerful finding: a nineteen percent reduction in deaths, just from a tax. Could you imagine trying to do that through education? You wouldn't be able to do it. It would cost way too much, and nobody would come up with that kinda money. Or you can just write a law that changes tax.

Now those are staggering findings, this these changes in behavior just from a tax. And it didn't come from small steps. It didn't come from advice like 'go get a dog and walk it.' That came from changing the law and placing a tax on the thing we want to discourage. And if a tax is done in this way, it potentially has many beneficial effects. So these different suggestions for food taxes have come up in countries, in England, in Ireland, in Australia. And it probably will happen at some point. So, the question I leave you with today is what role should government play in this whole process? And, is it taking a constructive role right now? That's for you to think about.

CONTENT-RICH VIDEOS IN L2 ACADEMIC LISTENING CONSTRUCT

7. We can infer that taxing fast food will _____
- (A) weaken public health
 - (B) raise people's objections
 - (C) make people wealthier
 - (D) increase fresh food sales
8. Cigarette tax rates are _____ across states in the US.
- (A) relatively similar
 - (B) largely different
 - (C) mostly high
 - (D) mostly low
9. In California, smoking-related deaths _____.
- (A) increased by 27%
 - (B) increased by 19%
 - (C) decreased by 27%
 - (D) decreased by 19%
10. We can infer that the teacher used the older data about tax rates to _____.
- (A) show that he is an expert
 - (B) present additional evidence
 - (C) compare historical data trends
 - (D) indicate an ineffective policy
11. Based on the listening, which statement is *NOT true*?
- (A) Educating about tobacco is better than taxing it.
 - (B) Tobacco taxes may fund anti-tobacco programs.
 - (C) Some countries have considered a food tax.
 - (D) Adding a new tax requires changing the law.
12. This lecture is mainly about _____.
- (A) tax rates and educational achievement
 - (B) tobacco tax rates across the US
 - (C) tobacco tax and anti-tobacco programs
 - (D) tax rates and human behavior

Testlet 3. Compassion (Questions 13-18)

Compassion is a really interesting thing to study because the world is full of more people who need help than we can possibly help. Right, if we try to feel compassion for everyone, it will be impossible and overwhelming. And so the question is: Out of all the people in the world who need help, how do we decide who it is most beneficial to help, ah who is most worthy of compassion? And what I wanna suggest to you is that one way that we go about deciding whether or not to help someone or whether or not to show compassion to them is based on a simple analysis: Do we see ourselves in them? And so I wanna suggest that one way compassion works is based on that simple metric, and that metric is similarity. The idea is: The more similar someone is to me, the more likely I am to feel compassion for them, even if they're suffering the same tragedy as another individual. And what this suggests is that distress is really in the eye of the beholder. How much compassion I feel for someone isn't a function of what's befallen them, it's a function of their links to me. Now if I said to you, on a battle field an American soldier comes upon a wounded member of Taliban and a wounded American soldier, and they feel more compassion towards the wounded American soldier, that might not be surprising to you. Those groups were in conflict for a long time. But what I wanna suggest is that this bias is so deeply embedded in the mind that we can see it even with the subtlest of cues.

And so the cues I really wanna look at, stripping it down to bare bones, is simple motor synchrony, right, moving in time together. If you move your body in time together, it's a marker that right now, in this moment, two individuals are one. Their purposes are joined, and their goals are joined. And those are the individuals who long-term are most likely going to help me. So, how do we do this? We bring individuals into a lab. We sit them down at a table, and they put on earphones. They think they're in the music perception study. And their goal is simple: Tap your hands to the tones you hear. The only difference is: Sometimes they tap their hands in unison, and sometimes the tones are random, so they tap in a completely asynchronous way. They don't talk, they don't do anything else. What happens next is that you see the partner who you were tapping with, engaging in another study that you're observing, in which they are being cheated by

another subject and being stuck with this onerous, tedious task. And then simply what we do is we ask them if they wanna help that person or not. We don't ask them as experimenters because that might add some extra pressure. Ah the end of the experiment, the computer simply says to them: There's more work to be done; if for some reason you'd like to help somebody else, please find one of the experimenters and let them know.

And what we've found, I have to admit to you, was rather astounding to me. The simple act of tapping your hands in time makes people feel more similar. Now they couldn't tell us why they were more similar, they would create stories about how they were similar. They didn't even talk to the other person, and yet they still felt similar. And what that similarity did is it gave the long-term mechanisms of the mind greater power to increase the compassion that we were gonna feel. And so the amount of compassion they felt was also influenced by whether or not they tapped in time with that person – if they did, they felt more compassion. But remember, in each case the person is victimized in the same way and cheated in exactly the same way. But how much compassion we feel for them is really a function of how similar we feel to them. Moreover, if you look at the decisions to help, there's a really large difference, right? 17 out of 35 people decided to help the person with whom they tapped their hands in time. Only 6 out of 34 decided to do that in cases where there was less similarity. And if you look at the time they spent helping, it's even more dramatic, right? If I feel similar to you, I helped you for much longer than I did if I felt that you and I were not similar.

CONTENT-RICH VIDEOS IN L2 ACADEMIC LISTENING CONSTRUCT

13. According to the speaker, we will probably feel more compassion for a person who _____.
- (A) is our soulmate or close relative
 - (B) got in serious trouble or difficulty
 - (C) suffers from the war's effects
 - (D) is similar to a famous celebrity
14. In the experiment, what happened after the tone tapping?
- (A) The tones were changed.
 - (B) One participant was cheated.
 - (C) Participants were seated.
 - (D) Experimenters helped participants.
15. If people tapped in time with a partner, they _____ their partners.
- (A) felt less similar to
 - (B) more often helped
 - (C) felt less compassion for
 - (D) more often looked at
16. Which statement is NOT true?
- (A) Moving together is a sign of having one goal.
 - (B) Participants were cheated in the same way.
 - (C) Participants knew why they felt similar.
 - (D) Talking was not allowed in the experiment.
17. Two partners would probably feel less similar if _____.
- (A) one of them was not cheated
 - (B) both of them were cheated
 - (C) their tasks were not tedious
 - (D) they heard tones at different times
18. The passage is mainly about _____ compassion.
- (A) what makes people feel
 - (B) how to do research on
 - (C) how to have people appreciate
 - (D) why it is important to study

Testlet 4. Exoplanets (Question 19-24)

This lecture focuses on one of the main methods for detecting exoplanets - the radial velocity method. As we'll discuss, the radial velocity method uses the motion, or the wobble, of a star to indicate the presence of a planet. As I alluded to when we talked about planetary motions, planets don't exactly orbit the Sun. We probably learned that the Sun's at the center and the planets orbit around the Sun. Well, that's not exactly true. Planets don't orbit the Sun. They orbit the barycenter, which is kind of a balance point. It's a balance point in mass between all the planets and the Sun. And that's hard to explain, when we consider all eight of the planets in our solar system. So let's just consider the biggest planet, Jupiter, and let's see how that goes with the Sun. So the Sun and the Jupiter play kind of cosmic balancing act. It's as if they're on a seesaw, if you will, and they have to balance each other. So if you put the Sun and Jupiter on a seesaw, Jupiter will be much farther away. It's 1,000 times less massive than the Sun. And the Sun will actually sit very close to the center, but not perfectly at the center. That balance point of the seesaw is what is called the barycenter. These two are balancing each other. So as Jupiter goes around in its orbit, the Sun also has to balance out Jupiter's mass and go round in its orbit. Turns out the barycenter of the Sun with respect to Jupiter is actually outside the surface of the Sun. And therefore, as Jupiter is going around in its orbit, the Sun, too, is going around in its orbit. So we can actually see, if you were looking at the solar system from above you'd actually see as Jupiter is going around, the Sun too is orbiting. It's making a much smaller orbit, but it too is making an orbit.

So this wobble, or this effect of a star having to orbit its own barycenter, is a telltale sign of planets around that star. But how can we detect them? There's some tricks that we can do for seeing the star's motion as it comes towards us and away from us. One of those tricks is the Doppler effect. The Doppler effect is an effect that most of you probably know because you've encountered it with sound. In fact, if you're walking down the street or you've heard a police car or an ambulance come towards you or going away from you, ah you hear, as that car comes towards you, the sound waves are compressed, and the pitch gets higher. Kind of goes -- beeeep. And as the car goes away from you, the sound waves are elongated, and the pitch goes down. Ah you hear kind of ahh baaooo. And of course, the engine or the siren of the police vehicle hasn't changed its pitch at all.

CONTENT-RICH VIDEOS IN L2 ACADEMIC LISTENING CONSTRUCT

It's just your perception. The waves have actually been compressed as they come to your ear. So many of us have heard that with sound. But the same principle applies to light. In fact, as an object comes towards you, the waves are compressed. The wavelength gets smaller, gets bluer. And as an object goes away from you, the waves are elongated, or get redder, as they get to longer wavelengths. And the faster an object moves, either towards you or away from you, the larger that shift is. So this is the light version of a Doppler effect.

But what can we use to study that? We know that now, if we can measure this light Doppler shift, if we can measure a star as it wobbles towards us, it should get a little bit bluer. And as it goes away from you, it should get a little bit redder. And in fact, that motion towards us and away from us is actually what's called radial velocity. That's why this technique is called radial velocity method. And we define radial velocity, positive radial velocity, as the motion away from us. So as the light gets a little bit redder, we call that positive radial velocity. As it gets bluer when it comes towards us, we call that negative radial velocity. So if we see that star go towards us, then away from us, then towards us, then away from us, we'll be detecting that star wobbling. And that's, again, the telltale sign that star has a planet in orbit. So what we can do is monitor these stars, take spectra, or distribution of colors coming from stars, and actually watch as these colors themselves wobble back and forth. We can actually observe the spectral features doing that, and the degree of the spectral shift tells us about the speed of that star's wobble. So the very first detection of an extrasolar planet around a star like our Sun was done in 1991 using this radial velocity method. It was done around the star 51 PEG. And so we call the exoplanet 51 PEG B, for the first exoplanet around that system.

CONTENT-RICH VIDEOS IN L2 ACADEMIC LISTENING CONSTRUCT

19. A barycenter is a/an _____.
- (A) planet's core or midpoint
 - (B) orbit or path of planets
 - (C) planet detection method
 - (D) balance point of planets
20. We can infer that a planet with less mass _____.
- (A) sits far from its barycenter
 - (B) has a smaller orbit
 - (C) has its barycenter inside
 - (D) completes its orbit faster
21. According to the speaker, which statement is *NOT true*?
- (A) The Sun goes around in its orbit.
 - (B) Car sirens change their pitch.
 - (C) Planets do not orbit the Sun.
 - (D) The Doppler Effect applies to light.
22. If an object comes *away from* us, it has _____.
- (A) longer waves
 - (B) higher pitch
 - (C) bluer colors
 - (D) negative radial velocity
23. What would be a sign that a planet is orbiting?
- (A) blue colors
 - (B) red colors
 - (C) both blue and red colors
 - (D) no colors and shorter waves
24. This passage is mainly about detecting the _____.
- (A) barycenter of a planet
 - (B) motion of a planet
 - (C) planets' sound waves
 - (D) orbits of Jupiter and the Sun

CONTENT-RICH VIDEOS IN L2 ACADEMIC LISTENING CONSTRUCT

Academic Listening Test Answer Key

Testlet 1. Homeostasis	Testlet 2. Food Tax	Testlet 3. Compassion	Testlet 4. Exoplanets
1. B	7. D	13. A	19. D
2. C	8. B	14. B	20. A
3. A	9. D	15. B	21. B
4. A	10. C	16. C	22. A
5. B	11. A	17. D	23. C
6. D	12. D	18. A	24. B

CONTENT-RICH VIDEOS IN L2 ACADEMIC LISTENING CONSTRUCT

Academic Listening Test – Table of Specification

Listening Testlets	Sub-constructs			# items	%
	Main Ideas	Details	Inferences		
Testlet 1. Homeostasis	1	3	2	6	25%
a) 03:58 b) 1 speaker c) Physical science d) moderately fast e) video-based version: 20.6% pictures, 40.0% graphs	6	1, 2, 5	3, 4		
Testlet 2. Food Tax	1	3	2	6	25%
a) 04:08 b) 1 speaker c) Social science d) moderately fast e) video-based version: 20.9% pictures, 39.7% graphs	12	8, 9, 11	7, 10		
Testlet 3. Compassion	1	3	2	6	25%
a) 03:57 b) 1 speaker c) Social science d) moderately fast e) video-based version: 17.1% pictures, 42.5% graphs	18	14, 15, 16	13, 17		
Testlet 4. Exoplanets	1	3	2	6	25%
a) 04:16 b) 1 speaker c) Physical Science d) moderately fast e) video-based version: 18.6% pictures, 40.7% graphs	24	19, 21, 22	20, 23		
Items per sub-construct	4	12	8	24	100%
Points per item	1	1	1		
Points per sub-construct	4	12	8	Raw Pts: 24	

Appendix E

Item Video-Dependence Survey

Video-Dependence Survey (AV) - Teacher

Consent

You are being invited to participate in a research study titled “*The Role of Content-Rich Videos in the L2 Academic Listening Assessment Construct.*” This study is being done by **Roman Lesnov** from Northern Arizona University. The purpose of this research study is to justify the use of visual information in second language tests of listening comprehension. If you agree to take part in the study, you will be asked to provide your judgement on the degree to which listening comprehension items in a test are answerable from the video input. It will take you approximately **1 hour** to complete.

Specifically, you will be asked to:

- Listen to or watch the four listening testlets
- Read each individual item within each testlet
- Provide your judgement using the scale developed by the researcher or provide your answers for each item

You may not directly benefit from this research; however, we hope that your participation in the study may help second language teachers know more about the ways to increase the effectiveness of teaching and testing listening.

We believe there are no known risks associated with this research study; however, as with any online related activity the risk of a breach of confidentiality is always possible. To the best of our ability your answers in this study will remain confidential. We will minimize any risks by keeping your responses confidential.

Your participation in this study is completely voluntary and you can withdraw at any time. You are free to skip any question that you choose. If you choose not to participate it will not affect your relationship with Northern Arizona University or result in any other penalty or loss of benefits to which you are otherwise entitled.

If you have questions about this project or if you have a research-related problem, you may contact the researcher(s), **Roman Lesnov, (1)929-225-9330**. If you have any questions concerning your rights as a research subject, you may contact Northern Arizona University IRB Office at irb@nau.edu or (928) 523-9551.

By clicking “NEXT,” I affirm that I am over 18 years of age and agree that the information may be used in the research project described above.

Instructions

Dear Teacher!

You will watch four lecture videos about (1) homeostasis, (2) taxes, (3) compassion, and (4) exoplanets. For each video, follow the directions below (you will go through each step automatically).

Directions:

1. Prior to watching, read comprehension questions.
2. Attentively watch the video. Do not pause the video at this time. Take notes if needed.
3. Choose the best answer for each comprehension question.
4. Use your judgement to answer the following questions. You can replay the video or its parts.

To what degree can the video-based visual cues help a test-taker to answer this question correctly?

Choose from 1 (not helpful) to 5 (very helpful).

Where is the answer to this question located in the video stream?

If applicable, indicate the time interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

Chose all that apply:

- pictures/photos;
- graphs/schemes/charts;
- text;
- the speaker's non-verbal cues

Understanding the scale

When asked about the degree of helpfulness of visual cues for answering questions, you will use a semantic differential scale ranging from 1 to 5. To help you understand this scale, think about the bipolar ends of the scale in the following way:

1 - visuals do not contain the answer to the question

5 - visuals clearly contain the answer to the question

Lecture 1. Homeostasis

Read the questions below. Do not answer the questions at this time.

1. According to the speaker, which statement about weight is *true*?

- Young people now have serious weight issues.
- We do not have to help our body control its weight.
- Eating food makes our weight quite unstable.
- We do not normally talk much about weight.

2. The normal body temperature range is _____ degrees Celsius.

- 36.5-37.0
- 36.0-37.5
- 36.5-37.5
- 37.5-38.5

3. We can infer that thermostats are _____.

- quite familiar to students
- weakly related to the lecture
- not helpful for understanding homeostasis
- in a perfect condition in college dorm rooms

4. Our body will most likely send the *second control signal* when _____.

- we have a fever
- we are cold
- our temperature is normal
- our temperature drops fast

5. Temperature control is important because it _____.

- slows harmful chemical reactions
- helps molecules work effectively
- increases the number of enzymes
- manages the body's feedback

6. This lecture is mainly about _____.

- how our body keeps its weight constant
- which body parameters are most important
- why body temperature is important
- how our body controls its environment

Lecture 1. Watch the lecture. You may take notes if needed.

Homeostasis. Question 1

Answer all the questions. You can replay the video or its parts.

1. According to the speaker, which statement about weight is *true*?*

- Young people now have serious weight issues.
- We do not have to help our body control its weight.
- Eating food makes our weight quite unstable.
- We do not normally talk much about weight.

To what degree can the video-based visual cues help a test-taker to answer this question correctly? Choose from 1 (not helpful) to 5 (very helpful).

Not helpful	1	2	3	4	5	Very helpful
--------------------	---	---	---	---	---	---------------------

Where is the answer to this question located in the video stream?

Indicate the interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

	Pictures photos	Graphs Schemes charts	text	the speaker's non- verbal cues
Check all that apply.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Homeostasis. Question 2

Answer all the questions. You can replay the video or its parts.

2. The normal body temperature range is _____ degrees Celsius.*

- 36.5-37.0
- 36.0-37.5
- 36.5-37.5
- 37.5-38.5

To what degree can the video-based visual cues help a test-taker to answer this question correctly? Choose from 1 (not helpful) to 5 (very helpful).

Not helpful	1	2	3	4	5	Very helpful
--------------------	---	---	---	---	---	---------------------

Where is the answer to this question located in the video stream?

Indicate the interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

	Pictures photos	Graphs Schemes charts	text	the speaker's non- verbal cues
Check all that apply.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Homeostasis. Question 3

Answer all the questions. You can replay the video or its parts.

3. We can infer that thermostats are _____.*

- quite familiar to students
- weakly related to the lecture
- not helpful for understanding homeostasis
- in a perfect condition in college dorm rooms

To what degree can the video-based visual cues help a test-taker to answer this question correctly? Choose from 1 (not helpful) to 5 (very helpful).

Not helpful 1 2 3 4 5 **Very helpful**

Where is the answer to this question located in the video stream?

Indicate the interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

	Pictures photos	Graphs Schemes charts	text	the speaker's non- verbal cues
Check all that apply.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Homeostasis. Question 4

Answer all the questions. You can replay the video or its parts.

4. Our body will most likely send the *second control signal* when _____.*

- we have a fever
- we are cold
- our temperature is normal
- our temperature drops fast

To what degree can the video-based visual cues help a test-taker to answer this question correctly? Choose from 1 (not helpful) to 5 (very helpful).

Not helpful 1 2 3 4 5 Very helpful

Where is the answer to this question located in the video stream?

Indicate the interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

	Pictures photos	Graphs Schemes charts	text	the speaker's non- verbal cues
Check all that apply.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Homeostasis. Question 5

Answer all the questions. You can replay the video or its parts.

5. Temperature control is important because it _____.*

- slows harmful chemical reactions
- helps molecules work effectively
- increases the number of enzymes
- manages the body's feedback

To what degree can the video-based visual cues help a test-taker to answer this question correctly? Choose from 1 (not helpful) to 5 (very helpful).

Not helpful	1	2	3	4	5	Very helpful
--------------------	---	---	---	---	---	---------------------

Where is the answer to this question located in the video stream?

Indicate the interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

	Pictures photos	Graphs Schemes charts	text	the speaker's non- verbal cues
Check all that apply.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Homeostasis. Question 6

Answer all the questions. You can replay the video or its parts.

6. This lecture is mainly about _____.*

- how our body keeps its weight constant
- which body parameters are most important
- why body temperature is important
- how our body controls its environment

To what degree can the video-based visual cues help a test-taker to answer this question correctly? Choose from 1 (not helpful) to 5 (very helpful).

Not helpful	1	2	3	4	5	Very helpful
--------------------	---	---	---	---	---	---------------------

Where is the answer to this question located in the video stream?

Indicate the interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

	Pictures photos	Graphs Schemes charts	text	the speaker's non- verbal cues
Check all that apply.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Lecture 2. Taxes

Read the questions below. Do not answer the questions at this time.

7. We can infer that taxing fast food will _____

- weaken public health
- raise people's objections
- make people wealthier
- increase fresh food sales

8. Cigarette tax rates are _____ across states in the US.

- relatively similar
- largely different
- mostly high
- mostly low

9. In California, smoking-related deaths _____.

- increased by 27%
- increased by 19%
- decreased by 27%
- decreased by 19%

10. We can infer that the teacher used the older data about tax rates to _____.

- show that he is an expert
- present additional evidence
- compare historical data trends
- indicate an ineffective policy

11. Based on the listening, which statement is *NOT true*?

- Educating about tobacco is better than taxing it.
- Tobacco taxes may fund anti-tobacco programs.
- Some countries have considered a food tax.
- Adding a new tax requires changing the law.

12. This lecture is mainly about _____.

- tax rates and educational achievement
- tobacco tax rates across the US
- tobacco tax and anti-tobacco programs
- tax rates and human behavior

Lecture 2. Watch the lecture. You may take notes if needed.

Taxes. Question 7

Answer all the questions. You can replay the video or its parts.

7. We can infer that taxing fast food will _____ *

- weaken public health
- raise people's objections
- make people wealthier
- increase fresh food sales

To what degree can the video-based visual cues help a test-taker to answer this question correctly? Choose from 1 (not helpful) to 5 (very helpful).

Not helpful	1	2	3	4	5	Very helpful
--------------------	---	---	---	---	---	---------------------

Where is the answer to this question located in the video stream?

Indicate the interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

	Pictures photos	Graphs Schemes charts	text	the speaker's non- verbal cues
Check all that apply.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Taxes. Question 8

Answer all the questions. You can replay the video or its parts.

8. Cigarette tax rates are _____ across states in the US.*

- relatively similar
- largely different
- mostly high
- mostly low

To what degree can the video-based visual cues help a test-taker to answer this question correctly? Choose from 1 (not helpful) to 5 (very helpful).

Not helpful	1	2	3	4	5	Very helpful
--------------------	---	---	---	---	---	---------------------

Where is the answer to this question located in the video stream?

Indicate the interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

	Pictures photos	Graphs Schemes charts	text	the speaker's non- verbal cues
Check all that apply.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Taxes. Question 9

Answer all the questions. You can replay the video or its parts.

9. In California, smoking-related deaths _____.*

- increased by 27%
- increased by 19%
- decreased by 27%
- decreased by 19%

To what degree can the video-based visual cues help a test-taker to answer this question correctly? Choose from 1 (not helpful) to 5 (very helpful).

Not helpful	1	2	3	4	5	Very helpful
--------------------	---	---	---	---	---	---------------------

Where is the answer to this question located in the video stream?

Indicate the interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

	Pictures photos	Graphs Schemes charts	text	the speaker's non- verbal cues
Check all that apply.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Taxes. Question 10

Answer all the questions. You can replay the video or its parts.

10. We can infer that the teacher used the older data about tax rates to _____.*

- show that he is an expert
- present additional evidence
- compare historical data trends
- indicate an ineffective policy

To what degree can the video-based visual cues help a test-taker to answer this question correctly? Choose from 1 (not helpful) to 5 (very helpful).

Not helpful	1	2	3	4	5	Very helpful
--------------------	---	---	---	---	---	---------------------

Where is the answer to this question located in the video stream?

Indicate the interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

	Pictures photos	Graphs Schemes charts	text	the speaker's non- verbal cues
Check all that apply.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Taxes. Question 11

Answer all the questions. You can replay the video or its parts.

11. Based on the listening, which statement is *NOT true*?*

- Educating about tobacco is better than taxing it.
- Tobacco taxes may fund anti-tobacco programs.
- Some countries have considered a food tax.
- Adding a new tax requires changing the law.

To what degree can the video-based visual cues help a test-taker to answer this question correctly? Choose from 1 (not helpful) to 5 (very helpful).

Not helpful	1	2	3	4	5	Very helpful
--------------------	---	---	---	---	---	---------------------

Where is the answer to this question located in the video stream?

Indicate the interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

	Pictures photos	Graphs Schemes charts	text	the speaker's non- verbal cues
Check all that apply.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Taxes. Question 12

Answer all the questions. You can replay the video or its parts.

12. This lecture is mainly about _____.*

- tax rates and educational achievement
- tobacco tax rates across the US
- tobacco tax and anti-tobacco programs
- tax rates and human behavior

To what degree can the video-based visual cues help a test-taker to answer this question correctly? Choose from 1 (not helpful) to 5 (very helpful).

Not helpful	1	2	3	4	5	Very helpful
--------------------	---	---	---	---	---	---------------------

Where is the answer to this question located in the video stream?

Indicate the interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

	Pictures photos	Graphs Schemes charts	text	the speaker's non- verbal cues
Check all that apply.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Lecture 3. Compassion

Read the questions below. Do not answer the questions at this time.

13. According to the speaker, we will probably feel more compassion for a person who _____.

- is our soulmate or close relative
- got in serious trouble or difficulty
- suffers from the war's effects
- is similar to a famous celebrity

14. In the experiment, what happened after the tone tapping?

- The tones were changed.
- One participant was cheated.
- Participants were seated.
- Experimenters helped participants.

15. If people tapped in time with a partner, they _____ their partners.

- felt less similar to
- more often helped
- felt less compassion for
- more often looked at

16. Which statement is *NOT* true?

- Moving together is a sign of having one goal.
- Participants were cheated in the same way.
- Participants knew why they felt similar.
- Talking was not allowed in the experiment.

17. Two partners would probably feel less similar if _____.

- one of them was not cheated
- both of them were cheated
- their tasks were not tedious
- they heard tones at different times

18. The passage is mainly about _____ compassion.

- what makes people feel
- how to do research on
- how to have people appreciate
- why it is important to study

Lecture 3. Watch the lecture. You may take notes if needed.

Compassion. Question 13

Watch the video and answer the questions. You can replay the video or its parts.

13. According to the speaker, we will probably feel more compassion for a person who _____.*

- is our soulmate or close relative
- got in serious trouble or difficulty
- suffers from the war's effects
- is similar to a famous celebrity

To what degree can the video-based visual cues help a test-taker to answer this question correctly? Choose from 1 (not helpful) to 5 (very helpful).

Not helpful 1 2 3 4 5 Very helpful

Where is the answer to this question located in the video stream?

Indicate the interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

	Pictures photos	Graphs Schemes charts	text	the speaker's non- verbal cues
Check all that apply.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Compassion. Question 14

Watch the video and answer the questions. You can replay the video or its parts.

14. In the experiment, what happened after the tone tapping?*

- The tones were changed.
- One participant was cheated.
- Participants were seated.
- Experimenters helped participants.

To what degree can the video-based visual cues help a test-taker to answer this question correctly? Choose from 1 (not helpful) to 5 (very helpful).

Not helpful 1 2 3 4 5 **Very helpful**

Where is the answer to this question located in the video stream?

Indicate the interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

	Pictures photos	Graphs Schemes charts	text	the speaker's non- verbal cues
Check all that apply.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Compassion. Question 15

Watch the video and answer the questions. You can replay the video or its parts.

15. If people tapped in time with a partner, they _____ their partners.*

- felt less similar to
- more often helped
- felt less compassion for
- more often looked at

To what degree can the video-based visual cues help a test-taker to answer this question correctly? Choose from 1 (not helpful) to 5 (very helpful).

Not helpful 1 2 3 4 5 Very helpful

Where is the answer to this question located in the video stream?

Indicate the interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

	Pictures photos	Graphs Schemes charts	text	the speaker's non- verbal cues
Check all that apply.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Compassion. Question 16

Watch the video and answer the questions. You can replay the video or its parts.

16. Which statement is *NOT* true?*

- Moving together is a sign of having one goal.
- Participants were cheated in the same way.
- Participants knew why they felt similar.
- Talking was not allowed in the experiment.

To what degree can the video-based visual cues help a test-taker to answer this question correctly? Choose from 1 (not helpful) to 5 (very helpful).

Not helpful	1	2	3	4	5	Very helpful
--------------------	---	---	---	---	---	---------------------

Where is the answer to this question located in the video stream?

Indicate the interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

	Pictures photos	Graphs Schemes charts	text	the speaker's non- verbal cues
Check all that apply.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Compassion. Question 17

Watch the video and answer the questions. You can replay the video or its parts.

17. Two partners would probably feel less similar if _____.*

- one of them was not cheated
- both of them were cheated
- their tasks were not tedious
- they heard tones at different times

To what degree can the video-based visual cues help a test-taker to answer this question correctly? Choose from 1 (not helpful) to 5 (very helpful).

Not helpful	1	2	3	4	5	Very helpful
--------------------	---	---	---	---	---	---------------------

Where is the answer to this question located in the video stream?

Indicate the interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

	Pictures photos	Graphs Schemes charts	text	the speaker's non- verbal cues
Check all that apply.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Compassion. Question 18

Watch the video and answer the questions. You can replay the video or its parts.

18. The passage is mainly about _____ compassion.*

- what makes people feel
- how to do research on
- how to have people appreciate
- why it is important to study

To what degree can the video-based visual cues help a test-taker to answer this question correctly? Choose from 1 (not helpful) to 5 (very helpful).

Not helpful	1	2	3	4	5	Very helpful
--------------------	---	---	---	---	---	---------------------

Where is the answer to this question located in the video stream?

Indicate the interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

	Pictures photos	Graphs Schemes charts	text	the speaker's non- verbal cues
Check all that apply.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Lecture 4. Exoplanets

Read the questions below. Do not answer the questions at this time.

19. A barycenter is a/an _____.

- planet's core or midpoint
- orbit or path of planets
- planet detection method
- balance point of planets

20. We can infer that a planet with less mass _____.

- sits far from its barycenter
- has a smaller orbit
- has its barycenter inside
- completes its orbit faster

21. According to the speaker, which statement is *NOT true*?

- The Sun goes around in its orbit.
- Car sirens change their pitch.
- Planets do not orbit the Sun.
- The Doppler Effect applies to light.

22. If an object comes *away from us*, it has _____.

- longer waves
- higher pitch
- bluer colors
- negative radial velocity

23. What would be a sign that a planet is orbiting?

- blue colors
- red colors
- both blue and red colors
- no colors and shorter waves

24. This passage is mainly about detecting the _____.

- barycenter of a planet
- motion of a planet
- planets' sound waves
- orbits of Jupiter and the Sun

Lecture 4. Watch the lecture. You may take notes if needed.

Exoplanets. Question 19

Answer all the questions. You can replay the video or its parts.

19. A barycenter is a/an _____.*

- planet's core or midpoint
- orbit or path of planets
- planet detection method
- balance point of planets

To what degree can the video-based visual cues help a test-taker to answer this question correctly? Choose from 1 (not helpful) to 5 (very helpful).

Not helpful	1	2	3	4	5	Very helpful
--------------------	---	---	---	---	---	---------------------

Where is the answer to this question located in the video stream?

Indicate the interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

	Pictures photos	Graphs Schemes charts	text	the speaker's non- verbal cues
Check all that apply.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Exoplanets. Question 20

Answer all the questions. You can replay the video or its parts.

20. We can infer that a planet with less mass _____.*

- sits far from its barycenter
- has a smaller orbit
- has its barycenter inside
- completes its orbit faster

To what degree can the video-based visual cues help a test-taker to answer this question correctly? Choose from 1 (not helpful) to 5 (very helpful).

Not helpful	1	2	3	4	5	Very helpful
--------------------	---	---	---	---	---	---------------------

Where is the answer to this question located in the video stream?

Indicate the interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

	Pictures photos	Graphs Schemes charts	text	the speaker's non- verbal cues
Check all that apply.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Exoplanets. Question 21

Answer all the questions. You can replay the video or its parts.

21. According to the speaker, which statement is *NOT true*?*

- The Sun goes around in its orbit.
- Car sirens change their pitch.
- Planets do not orbit the Sun.
- The Doppler Effect applies to light.

To what degree can the video-based visual cues help a test-taker to answer this question correctly? Choose from 1 (not helpful) to 5 (very helpful).

Not helpful	1	2	3	4	5	Very helpful
--------------------	---	---	---	---	---	---------------------

Where is the answer to this question located in the video stream?

Indicate the interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

	Pictures photos	Graphs Schemes charts	text	the speaker's non- verbal cues
Check all that apply.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Exoplanets. Question 22

Answer all the questions. You can replay the video or its parts.

22. If an object comes *away from us*, it has _____.*

- longer waves
- higher pitch
- bluer colors
- negative radial velocity

To what degree can the video-based visual cues help a test-taker to answer this question correctly? Choose from 1 (not helpful) to 5 (very helpful).

Not helpful	1	2	3	4	5	Very helpful
--------------------	---	---	---	---	---	---------------------

Where is the answer to this question located in the video stream?

Indicate the interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

	Pictures photos	Graphs Schemes charts	text	the speaker's non- verbal cues
Check all that apply.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Exoplanets. Question 23

Answer all the questions. You can replay the video or its parts.

23. What would be a sign that a planet is orbiting?*

- blue colors
- red colors
- both blue and red colors
- no colors and shorter waves

To what degree can the video-based visual cues help a test-taker to answer this question correctly? Choose from 1 (not helpful) to 5 (very helpful).

Not helpful	1	2	3	4	5	Very helpful
--------------------	---	---	---	---	---	---------------------

Where is the answer to this question located in the video stream?

Indicate the interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

	Pictures photos	Graphs Schemes charts	text	the speaker's non- verbal cues
Check all that apply.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Exoplanets. Question 24

Answer all the questions. You can replay the video or its parts.

24. This passage is mainly about detecting the _____.*

- barycenter of a planet
- motion of a planet
- planets' sound waves
- orbits of Jupiter and the Sun

To what degree can the video-based visual cues help a test-taker to answer this question correctly? Choose from 1 (not helpful) to 5 (very helpful).

Not helpful	1	2	3	4	5	Very helpful
--------------------	---	---	---	---	---	---------------------

Where is the answer to this question located in the video stream?

Indicate the interval that contains or alludes to the answer (e.g., 3:03-3:34).

What type(s) of visuals will help a test-taker to obtain the answer?

	Pictures photos	Graphs Schemes charts	text	the speaker's non- verbal cues
Check all that apply.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Last Questions

**Did you experience any problems with technology while taking the test?
(Slow internet connection? Glitches? Slow videos?)**

What is your email address?

Thank You!

THANK YOU!

**Thank you for taking our test. Your response is very
important to us.**

Appendix F

Classifying ALC Items as Video-Dependent vs. Video-Independent

Table F.1

Collective Data for Items with Originally Video-Dependent Design (k = 16)

Lecture	Item #	Item type	Video-dependence survey		Muted ALC video-based test		Final decision
			Teachers' ratings	Learners' ratings	Teachers' total	Learners' total	
Homeostasis	Item 2	local	5.00	4.33	1**	3	YES
	Item 4	global	3.67	4.33	3	2	YES
	Item 5	local	3.67	1.00*	2	0**	YES
Food Tax	Item 6	global	2.33*	2.67*	0**	1**	NO
	Item 7	global	2.33*	2.67*	2	2	YES
	Item 8	local	4.67	5.00	3	3	YES
	Item 9	local	4.33	5.00	3	3	YES
Compassion	Item 12	global	2.67*	3.33	2	2	YES
	Item 13	global	3.67	2.67*	2	3	YES
	Item 14	local	1.00*	2.00*	0**	0**	NO
	Item 15	local	4.00	4.67	2	2	YES
Exoplanets	Item 18	global	2.00*	2.67*	3	2	YES
	Item 19	local	4.00	2.67*	2	3	YES
	Item 20	global	2.67*	4.33	3	3	YES
	Item 22	local	4.00	4.33	3	3	YES
	Item 24	global	2.00*	2.00*	0**	1**	NO
<i>M</i>			3.25	3.35	1.93	2.06	
<i>SD</i>			1.13	1.23	1.12	1.06	

Note: * = lower than or equal to the 3.00 cut-off; ** = lower than the cut-off of 2; YES = confirmed as video-dependent; NO = re-classified as video-independent

Table F.2

Collective Data for Items with Originally Video-Independent Design (k = 8)

Lecture	Item #	Item type	Video-dependence survey		Muted ALC video-based test		Final decision
			Teachers' ratings	Learners' ratings	Teachers' total	Learners' total	
Homeostasis	Item 1	local	1.00	1.33	0	0	NO
	Item 3	global	2.00	2.33	1	1	NO
Food Tax	Item 10	global	2.33	2.33	2**	1	NO
	Item 11	local	1.33	1.00	2**	0	NO
Compassion	Item 16	local	1.00	1.33	0	1	NO
	Item 17	global	2.67	3.00	2**	0	NO
Exoplanets	Item 21	local	1.33	2.67	1	0	NO
	Item 23	global	2.33	4.00*	3**	2**	YES
<i>M</i>			1.75	2.25	1.38	0.63	
<i>SD</i>			0.66	1.01	1.06	0.74	

Note: * = higher than the 3.00 cut-off; ** = higher than the cut-off of 2; YES = re-classified as video-dependent; NO = confirmed as video-independent

Appendix G

YouTube Video Listening Passages for the Anchor Test

Title (Citation)	Lecture type	Link / Time boundaries / University affiliation
Cybersecurity (University of Delaware, 2015)	traditional	https://www.youtube.com/watch?v=3MkFO6EALi8 17:45 – 22:00 University of Delaware, USA
Language (Hilpert, 2014)	online	https://www.youtube.com/watch?v=up0yVJWf9zQ 0:00 – 3:47 University of Neuchâtel, Switzerland

Appendix H

Anchor Listening Test (scripts, items, table of specifications)

Anchor testlet 1. Cybersecurity (Questions 1-6).

Testlet 1. Cybersecurity

The reality is when you are online there is no way to be sure that the person you think you are communicating with or the website you're ... are going to is really that person or that website. There is no 100% certainty with the basic architecture of the network. And so ... we've had to think about how do you manage this problem. The problem is maybe best encapsulated by a New York ... New Yorker magazine cartoon – I think it goes back fifteen to twenty years – it's back in the days of big clunky ah PCs on the desk. And there's a drawing of a PC and there's two dogs talking to each other. And one dog says to the other: "On the Internet nobody knows you're a dog." And in many ways that sums up the problem. So with this lack of trust and with the ability of people to masquerade as others and use it as a way to gain entry to our own networks, what we've seen again and again is the capability that people have if they're bad actors to corrupt information, to steal information, to deny access or introduce latency or delay in the transmission of information, to destroy and overwhelm networks and of course to steal all kinds of information for financial gain.

If I would group ... these types of consequences, I would say in the main they fall into three main categories. The one that's maybe the most long-standing set of security challenges and the one that we still read about the most and probably the one that touches us personally the most is the use of the network to steal financial information for the purposes of committing fraud – identity information, credit card information, access to bank accounts. Ah as you've read there've been literally millions of dollars stolen in this way. In the last couple of years, for example, there was one organized criminal effort to gain access to ATMs. What they did was they hacked into a couple of firms overseas, they were managing debit cards and ATM withdrawal cards, and they had the withdrawal limits on those cards removed. Then, on a single day, ah individuals working as part of this conspiracy were sent out to ATM machines all over the world to withdraw all the

money from the machines. Because the withdrawal limits were gone, they could take every cash bit of cash that was in those machines. And on a single day before it was shut down tens of millions of dollars were stolen. So, that's a classic example of the fact that because the Internet is now where the money is, it's like [??]. To paraphrase [??], you don't have to rob banks any more by going in with a gun – you just rob it through the ATM or the credit card.

A second area of things that we have seen are denial of service attacks. Ah these aren't maybe the most sophisticated attacks, they don't ultimately destroy ah systems or networks, they don't kill people, but they interfere with the ability to get access to your ... perhaps your bank or some other facility that you need to communicate with. And they create an enormous burden and dragging expense for enterprises.

But the third and most consequential from a national security standpoint, the third type of category of attacks we worry about are attacks that actually could be corruptive or destructive. Imagine what would happen if ah malevolent actors penetrated into banks and were able able over a period of time, in a very subtle way, to change bank records. If you didn't have a back-up for transactions, you might have a crisis of confidence in banks something like what we saw in 2008 when we had our financial crisis. You could have destruction of critical infrastructure but unlike in Sony which destroyed business enterprises, tools and and and information technology architecture, you could actually have attacks on critical infrastructure that deals with transportation – the train that I came up with, the airplane I'm flying, maybe power. And that could actually cause loss of life as well as significant economical property damage.

1. **The cartoon about two dogs was discussed to illustrate the ____.**
 - (A) solution for the lack of trust online
 - (B) disadvantages of early computers
 - (C) problem of trusting online resources
 - (C) types of online communication
2. **According to the speaker, which cyber-crime will touch people personally the most?**
 - (A) Stealing credit card information.
 - (B) Robbing a bank's ATM machine.
 - (C) Destroying a government office.
 - (D) Denying access to a bank website.
3. **To take all the money from ATMs, the criminals ____.**
 - (A) shut down power in the banks
 - (B) removed limits from credit cards
 - (C) robbed banks with a gun
 - (D) broke open the ATM machines
4. **For national security, the most serious category of cyber-crimes is ____.**
 - (A) stealing financial information
 - (B) corruptive or destructive attacks
 - (C) denial-of-service attacks
 - (D) robbing banks with a gun
5. **A cyber-attack of *the third type* would most likely target a ____.**
 - (A) family-owned business
 - (B) person's Facebook account
 - (C) government official's email
 - (D) country's energy system
6. **This lecture is mainly about the ____.**
 - (A) security problems of online systems
 - (B) secure access to bank computers
 - (C) problems of insecure ATM machines
 - (D) lack of trust among modern people

Anchor testlet 2. Language (Questions 7-12)

Hello there and welcome back to the introduction to English linguistics. In this video I'd like to talk about language acquisition - how do children learn a first language. And to start out with, let me give you a few basic facts about language learning. First of all, there is no genetic predisposition for learning any one particular language. A baby born to English-speaking parents will of course learn English but the same baby, if it grows up around people talking in Finnish or in Mandarin or in Sinhalese or in Welsh, will acquire any of those languages with the same speed and ease. All human languages are equally easy to acquire as a first language and not only that - children can acquire two or more first languages with ease. Yeah. Ahh having two or more first languages – that's called bilingualism or multilingualism and it has been shown that there are strong cognitive advantages to being bilingual. Bilinguals they have two language systems in their mind and in order to use one, they have to inhibit the other, so they have to concentrate on one thing and defocus another thing. And you can imagine that this helps in a whole lot of other cognitive tasks – you concentrate on one thing and selectively ignore the other thing.

Right. More facts about language acquisition. Ahhm I said that the process seems to be effortless - very easy and very rapid so that all essential parts of language – the grammatical structures, pronunciations, all of that, is in place by age five to six, so there kids talk pretty much like adults. Now of course they don't talk completely like adults – they don't have the same capabilities that adults have. Think of telling a good joke or understanding irony. There kids catch up over the years, but in terms of grammatical rules, pronunciations, knowledge of different words – the basics really are in place by age five to six. All this happens without formal instruction. You don't have to tell kids: this is right, this is wrong, this is what the rules are. They figure that out by themselves and, interestingly, the outcome is almost always the same. Everybody learns how to talk and ah even though there may be some people that talk really really well, that are super eloquent, that know how to talk in public, ahhm ... well this is a skill that you have to learn as an adult. Yeah ahh everybody learns instinctively how to talk well enough to hold a conversation.

CONTENT-RICH VIDEOS IN L2 ACADEMIC LISTENING CONSTRUCT

Right. Now there are certain puzzles associated with language acquisition. For one thing, kids say things that they've never heard before. How do they do that? Kids get things right without being corrected. How is that? How do they figure that out? And then they master grammar by age 5 but they don't master things that are equally complex or comparable to language like mathematics, differential equations. Mmm they have trouble doing that at age 15 and yet at age five they chatter away, yeah, they have trouble tying their shoelaces but they use relative clauses – that seems to be remarkable. Now linguists try to explain these puzzles with theories of language acquisition.

7. According to the speaker, which statement is *true*?

- (A) Children learn some languages faster than others.
- (B) Learning two languages may be difficult for children.
- (C) Children learn any human language equally easily.
- (D) Some children are slower at learning languages.

8. According to the speaker, bilingual children _____.

- (A) use two language systems at the same time
- (B) may have problems with concentrating
- (C) focus on one of the language systems
- (D) select a language that they know better

9. Children will most likely _____ by age 6.

- (A) need instruction to speak well
- (B) be able to tell many good jokes
- (C) be able to hold a conversation
- (D) know how to speak in public

10. Linguists hope to explain how children can _____.

- (A) say what they heard before
- (B) solve mathematical problems
- (C) understand language theories
- (D) speak right without correction

11. The lecture is mainly about _____ children.

- (A) formal language instruction for
- (B) learning a first language by
- (C) the facts about public speaking by
- (D) learning a foreign language by

12. The teacher will most likely talk next about _____.

- (A) the lives of famous linguists
- (B) what languages children should learn
- (C) how children learn a first language
- (D) how to teach children a first language

CONTENT-RICH VIDEOS IN L2 ACADEMIC LISTENING CONSTRUCT

Anchor Test Answer Key

Anchor testlet 1. Cybersecurity	Anchor testlet 2. Language
1. C	7. C
2. A	8. C
3. B	9. C
4. B	10. D
5. D	11. B
6. A	12. C

Anchor Test – Table of Specification

Listening Testlets	Sub-constructs			# items	%
	Main Ideas	Details	Inferences		
Testlet 1. Cybersecurity	1	3	2	6	50%
a) 04:15 b) 1 speaker c) Social science d) moderate speed	6	2, 3, 4	1, 5		
Testlet 2. Language	1	3	2	6	50%
a) 03:47 b) 1 speaker c) Social science d) slow to moderate speed	11	7, 8, 10	9, 12		
Items per subconstruct	2	6	4	12	100%
Points per item	1	1	1		
Points per subconstruct	2	6	4	Raw Pts: 12	

Appendix I

Test-takers' Questionnaire

Test-takers' Questionnaire. Audio-Only Version.

Section 1

1. How interesting was this lecture?

1	2	3	4	5	6
very boring					very interesting

2. How difficult was this lecture?

1	2	3	4	5	6
very easy					very difficult

3. How realistic was this lecture?

1	2	3	4	5	6
not realistic					very realistic

Section 2

4. Academic listening tests should have videos.

<input type="checkbox"/> Strongly Disagree	<input type="checkbox"/> Disagree	<input type="checkbox"/> Somewhat Disagree	<input type="checkbox"/> Somewhat Agree	<input type="checkbox"/> Agree	<input type="checkbox"/> Strongly Agree
--	-----------------------------------	--	---	--------------------------------	---

5. Academic listening tests should be audio-only.

<input type="checkbox"/> Strongly Disagree	<input type="checkbox"/> Disagree	<input type="checkbox"/> Somewhat Disagree	<input type="checkbox"/> Somewhat Agree	<input type="checkbox"/> Agree	<input type="checkbox"/> Strongly Agree
--	-----------------------------------	--	---	--------------------------------	---

6. With videos, academic listening tests are more valid.

- Strongly Disagree Disagree Somewhat Disagree Somewhat Agree Agree Strongly Agree

Section 3

7. What is your first language? _____

8. Which school are you in?

- Program in Intensive English, Northern Arizona University, USA
- English Language Center, Rochester Institute of Technology, USA
- Universidad de Sonora, Mexico
- Zaoksky Christian Institute of Arts and Sciences, Russia
- White Rabbit, Russia
- EnglishDom, the Russian Federation
- Skyeng, the Russian Federation
- Other

9. How old are you? _____

10. What is your gender?

- Male
- Female
- Other

Test-takers' Questionnaire. Video-Based Version

Section 1

A. How much of the video did you watch?

- I did **not** watch
- Little**
- About **half** of the video
- Most** of the video
- All** of the video

1. How interesting was this lecture?

1	2	3	4	5	6
not					very
realistic					realistic

2. How difficult was this lecture?

1	2	3	4	5	6
not					very
realistic					realistic

3. How realistic was this lecture?

1	2	3	4	5	6
not					very
realistic					realistic

B. Do you agree that you were able to answer some questions because you saw pictures and graphs?

- Strongly Disagree
- Disagree
- Somewhat Disagree
- Somewhat Agree
- Agree
- Strongly Agree

Section 2

4. Academic listening tests should have videos.

- Strongly Disagree Disagree Somewhat Disagree Somewhat Agree Agree Strongly Agree

5. Academic listening tests should be audio-only.

- Strongly Disagree Disagree Somewhat Disagree Somewhat Agree Agree Strongly Agree

6. With videos, academic listening tests are more valid.

- Strongly Disagree Disagree Somewhat Disagree Somewhat Agree Agree Strongly Agree

Section 3

7. What is your first language? _____

8. Which school are you in?

- Program in Intensive English, Northern Arizona University, USA
- English Language Center, Rochester Institute of Technology, USA
- Universidad de Sonora, Mexico
- Zaoksky Christian Institute of Arts and Sciences, Russia
- White Rabbit, Russia
- EnglishDom, the Russian Federation
- Skyeng, the Russian Federation
- Other

9. How old are you? _____

10. What is your gender?

- Male
- Female
- Other

CONTENT-RICH VIDEOS IN L2 ACADEMIC LISTENING CONSTRUCT

Table of Specifications for Test-takers' Questionnaire

Version	Content area (Construct)							Total
	Viewing behavior	Video effects on			Video helpfulness for answering questions	Use of videos in academic listening tests	Demographics	
		listening difficulty	motivation	authenticity				
Audio-only version		1 (#2)	1 (#1)	1 (#3)		3 (#4-6)	4 (#7-10)	10
		10%	10%	10%		30%	40%	100%
Video-based version	1 (#A)	1 (#2)	1 (#1)	1 (#3)	1 (#B)	3 (#4-6)	4 (#7-10)	12
	8.3%	8.3%	8.3%	8.3%	8.3%	25%	33.3%	100%

Appendix J

Teachers' Questionnaire

Is Seeing a Part of Academic Listening?

Page 1. Hello!

Hello! Thank you for your willingness to take the survey.

Please remember that the survey is **NOT** smartphone-friendly and **NOT** tablet-friendly. You can **only** take the survey **on a computer or a laptop.**

*You will be asked to enter your email address after you finish the questionnaire. This email address will be drawn into a raffle to win one of several \$40 prizes. If you wish to decline your participation but still enter into the drawing, please write to us at rlor84@gmail.com.

Page 2. Consent

You are being invited to participate in a research study titled “*The Role of Content-Rich Videos in the L2 Academic Listening Assessment Construct.*” This study is being done by **Roman Lesnov** from Northern Arizona University.

The purpose of this research study is to know more about the *opinions of English as a second language (ESL) teachers on the role of visual information in the listening comprehension process.* If you agree to take part in this study, you will be asked to complete an online questionnaire. This questionnaire will ask about ***how visual information affect listening comprehension,*** and it will take you approximately ***10 minutes*** to complete.

You may not directly benefit from this research; however, we hope that your participation in the study may help second language teachers know more about the ways to increase the effectiveness of teaching and testing listening.

We believe there are no known risks associated with this research study; however, as with any online related activity the risk of a breach of confidentiality is always possible. To the best of our ability your answers in this study will remain confidential. We will minimize any risks by keeping your responses confidential.

Your participation in this study is completely voluntary and you can withdraw at any time. You are free to skip any question that you choose. If you choose not to participate it not affect your relationship with Northern Arizona University or result in any other penalty or less of benefits to which you are otherwise entitled.

If you have questions about this project or if you have a research-related problem, you may contact the researcher(s), **Roman Lesnov, (1) 929-225-9330** If you have any questions concerning your rights as a research subject, you may contact Northern Arizona University IRB Office at irb@nau.edu or (928) 523-9551.

By clicking “NEXT,” I affirm that I am over 18 years of age and agree that the information may be used in the research project described above.

Page 3. Instructions

Think about **a typical university lecture.**

Now you will listen to a part of an academic lecture about cigarette tax. This lecture was created for a high-stakes academic English test like TOEFL, IELTS, etc.

When you are ready, press "NEXT."

Page 4. Listen to the Audio

LECTURE AUDIO

Page 5. Instructions

Now you will **watch** the video of the same lecture excerpt.

When you are ready, press "NEXT."

Page 6. Watch the video

LECTURE VIDEO

Page 7. To what extent do you agree with the following statements?

Select one of the six options for each statement.

	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
1. This video helps listeners understand what they hear.						
2. This video makes this listening more engaging for students.						
3. In real life, academic listeners normally process visual information similar to this video.						
4. Large-scale academic English listening tests should have videos like this.						
5. This video hinders academic listening comprehension.						
6. With this video, the lecture is more interesting.						
7. This video makes this listening less authentic.						
8. In a large-scale academic English listening test, this lecture should be video-based rather than audio-only.						
9. The video makes this listening boring.						
10. The video makes this listening more realistic.						
11. Videos like this should be used in large-scale academic English tests.						
12. This video makes listening comprehension easier.						
13. The video makes this listening more authentic.						
14. High-stakes academic English listening tests should NOT have videos like this.						
15. This video facilitates understanding of the listening message.						
16. The presence of this video increases interest in the message.						

Page 8. Last Questions

17. Which of the following best describes you?

second or foreign language teacher

assessment specialist (coordinator, proctor, rater, developer, etc.)

18. What is the highest level of education you've achieved in the area of second language teaching?

Teaching Certificate

Bachelor's

Master's

Doctorate

19. For how many years have you been teaching or working in a related area (creating tests, educating language teachers, etc.)?

20. What is your first (native) language?

21. What is your gender?

female

male

other

22. How old are you?

23. What is your email address? We will email you instructions on how to get 40\$, if you are a winner. If you prefer not to participate in the drawing, please ignore this question.

Thank You Page: Thank You!

Thank You!

Thank you for taking our survey. Your response is very important to us! If you win \$40, we will contact you using the email that you provided.

Teachers' Questionnaire: Table of Specifications

Item format	Item type	Items by content area (construct)					# items	
		Role of visuals				Background information		
		(1) Effects on difficulty	(2) Effects on motivation	(3) Effects on authenticity	(4) Use in listening tests			
Reduced (4-point) Likert Scale	Positive	1, 12, 15	2, 6, 16	3, 10, 13	4, 8, 11		12	16
	Negative	5	9	7	14		4	
Multiple Choice						17, 18, 19	3	7
Open-ended						20, 21, 22, 23	4	
		4	4	4	4	7		23

Appendix K

Rasch Analysis Specifications Template

```
; Example of a specification file

Title = Rasch analysis for RQ 1.1
facets = 6 ; six facets
arrange = m,N ; arrange tables by measure-descending, element number-ascending
positive = 1 ; for test-takers: greater score - greater measure
Noncenter = 1 ; test-takers facet is non-centered
pt-biserial = measure ; point-measure correlation
Null=999;
models=?B,?,?B,?,?,D ; interaction b/w delivery mode and video-dependence
models=?B,?,?B,?,?,D ; b/w delivery mode and proficiency
models=?B,?,?,?B,?,?,D ; b/w delivery mode and item type
models=?B,?,?B,?,?B,?,?,D ; b/w delivery mode, video-dependence, & proficiency
models=?B,?,?B,?,?B,?,?,D ; b/w delivery mode, video-dependence, & item type
models=?B,?,?B,?,?B,?,?B,?,D ; b/w delivery mode, video-dependence, proficiency, & item type
*
labels =
1,Test-takers
1-120
*
2,Mode
1=audio-only
2=video
*
3, Proficiency, D ; dummy facet
1=Lower
2=Higher
*
4, Item video-dependence
1=video-dependent
2=video-independent
*
5, Item type,
1=local,
2=global
*
6, Items ; multiple-choice items
1-24
*
data=
01,1,1,1,1,1,0 ;person 1, audio-only, higher prof, video-dep, local, item 1, wrong
01,1,1,2,1,2,1 ;person 1, audio-only, higher prof, video-ind, local, item 2, right
01,1,1,1,2,3,1 ;person 1, audio-only, higher prof, video-dep, global, item 3, right
....
01,1,1,2,2,24,0 ;person 1, audio-only, higher prof, video-ind, global, item 24, wrong
....
....
61,2,2,1,1,1,1 ;person 61, video-based, lower prof, video-dep, local, item 1, right
61,2,2,2,1,2,0 ;person 61, video-based, lower prof, video-ind, local, item 2, wrong
61,2,2,1,2,3,0 ;person 61, video-based, lower prof, video-dep, global, item 3, wrong
....
61,2,2,2,2,24,0 ;person 61, video-based, lower prof, video-ind, global, item 24, wrong
....
....
143,...
```

plus Repeaters' data (24 out of the 143 test-takers' data from taking the test in the opposite mode).

Appendix L

Post hoc Comparisons for Education-Experience Interaction (RQ 2.2)

Pairwise Comparisons						95% Confidence Interval for Differenced	
Experience	(I) Education	(J) Education	Mean Difference (I-J)	Std. Error	Significance	Lower Bound	Upper Bound
1-5 years	1	2	-2.908	1.602	.071	-6.064	0.247
		3	-2.964	1.502	.050	-5.923	-0.006
		4	-5.333	2.151	.014	-9.572	-1.095
	2	1	2.908	1.602	.071	-0.247	6.064
		3	-0.056	1.231	.964	-2.480	2.368
		4	-2.425	1.972	.220	-6.310	1.460
	3	1	2.964	1.502	.050	0.006	5.923
		2	0.056	1.231	.964	-2.368	2.480
		4	-2.369	1.891	.212	-6.095	1.357
	4	1	5.333	2.151	.014	1.095	9.572
		2	2.425	1.972	.220	-1.460	6.310
		3	2.369	1.891	.212	-1.357	6.095
6-10 years	1	2	-4.667	2.012	.021	-8.631	-0.702
		3	-4.560	1.721	.009	-7.950	-1.169
		4	-1.333	2.202	.545	-5.672	3.005
	2	1	4.667	2.012	.021	0.702	8.631
		3	0.107	1.312	.935	-2.478	2.692
		4	3.333	1.900	.081	-0.409	7.076
	3	1	4.560	1.721	.009	1.169	7.950
		2	-0.107	1.312	.935	-2.692	2.478
		4	3.226	1.588	.043	0.098	6.354
	4	1	1.333	2.202	.545	-3.005	5.672
		2	-3.333	1.900	.081	-7.076	0.409
		3	-3.226	1.588	.043	-6.354	-0.098
11-15 years	1	2	-1.333	1.563	.394	-4.413	1.746
		3	-0.501	1.364	.714	-3.188	2.185
		4	-1.056	1.883	.576	-4.765	2.654
	2	1	1.333	1.563	.394	-1.746	4.413
		3	0.832	1.152	.471	-1.437	3.101
		4	0.278	1.736	.873	-3.142	3.697
	3	1	0.501	1.364	.714	-2.185	3.188
		2	-0.832	1.152	.471	-3.101	1.437
		4	-0.554	1.558	.722	-3.625	2.516
	4	1	1.056	1.883	.576	-2.654	4.765
		2	-0.278	1.736	.873	-3.697	3.142
		3	0.554	1.558	.722	-2.516	3.625
16-20 years	1	2	-0.944	2.028	.642	-4.941	3.052
		3	1.597	1.397	.254	-1.155	4.350
		4	1.972	1.671	.239	-1.320	5.264
	2	1	0.944	2.028	.642	-3.052	4.941
		3	2.542	1.768	.152	-0.942	6.026
		4	2.917	1.992	.144	-1.007	6.841
	3	1	-1.597	1.397	.254	-4.350	1.155
		2	-2.542	1.768	.152	-6.026	0.942
		4	0.375	1.343	.780	-2.272	3.022
	4	1	-1.972	1.671	.239	-5.264	1.320
		2	-2.917	1.992	.144	-6.841	1.007
		3	-0.375	1.343	.780	-3.022	2.272
> 20 years	1	2	-1.187	2.002	.554	-5.132	2.757
		3	-1.973	1.718	.252	-5.357	1.412
		4	-2.818	1.745	.108	-6.256	0.619
	2	1	1.187	2.002	.554	-2.757	5.132
		3	-.785	1.292	.544	-3.331	1.761
		4	-1.631	1.328	.221	-4.247	0.985
	3	1	1.973	1.718	.252	-1.412	5.357
		2	0.785	1.292	.544	-1.761	3.331
		4	-0.846	.840	.315	-2.500	0.809
	4	1	2.818	1.745	.108	-0.619	6.256
		2	1.631	1.328	.221	-0.985	4.247
		3	0.846	.840	.315	-0.809	2.500

Note: Education: 1 – Certificate, 2 – Bachelor's, 3 – Master's, 4 – Doctorate; Bonferroni-adjusted $\alpha = .05/48 = .001$